High Dimensional Forecast Combinations Under Latent Structures

Zhentao Shi, Liangjun Su and Tian Xie*

June 12, 2020

Abstract

This paper presents a novel high dimensional forecast combination algorithm in the presence of many forecasts and potential latent group structures. In comparison with the standard estimation of the optimal weights that minimize the mean squared forecast error (MSFE), we propose to minimize the squared ℓ_2-norm of the weight vector subject to a relaxed version of the first-order conditions, yielding the ℓ_2-relaxation problem. Interestingly, it incorporate both the simple average (equal-weight) strategy and the conventional optimal weighting scheme as special cases by setting the tuning parameter to be sufficiently large or zero. A proper choice of the tuning parameter can achieve the bias and variance trade-off and deliver combined forecasts with roughly equal groupwise weights when the variance-covariance (VC) matrix of the individual forecast errors exhibit latent group structures. This is consistent with the intuition that one should assign the same weights to all individuals within the same group and potentially distinct weights to individuals in different groups when the forecast error VC matrix exactly exhibits a block equicorrelation structure. When the VC matrix is contaminated with a noisy component, we show that the resultant ℓ_2-relaxed weights approximate to the optimal infeasible groupwise equal weights. As the number of forecasters N and time periods T go to infinity and N may potentially be much larger than T, the asymptotic optimality is established by exploiting the duality between the sup-norm restriction and the high-dimensional sparse ℓ_1-norm penalization. Both simulated and real data are used to evaluate the finite sample performance of our proposed estimator in comparison with some competitive methods, and they demonstrate the superb performance of our method.

Key Words: Factor models; forecast combination puzzle; high dimension; Lasso; latent group; machine learning; optimality.

JEL Classification: C22, C53, C55

*Zhentao Shi: zhentao.shi@cuhk.edu.hk, Department of Economics, 928 Esther Lee Building, the Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China. Tel: (852) 3943-1432. Fax (852) 2603-5805.
Zhentao Shi acknowledges financial support from the Research Grants Council (RGC) No.14500118.
Liangjun Su: School of Economics and Management, Tsinghua University.
Tian Xie: School of Business, Shanghai University of Fiance and Economics

1 Introduction

Since the seminal work of Bates and Granger (1969), forecast combination has become very popular in the forecasting literature. For an excellent review, see Timmermann (2006). In this paper we consider high-dimensional forecast combinations in the presence of many forecasts. Suppose that y_{t+1} is an outcome variable of interest and there are N forecasts, $\{f_{it}\}_{i\in[N]}$, available at time t for y_{t+1}, where $t \in [T] \equiv \{1, 2, ..., T\}$ and $[N] \equiv \{1, 2, ..., N\}$. Let $\mathbf{f}_t = (f_{1t}, ..., f_{Nt})'$. The interest is to find an $N \times 1$ weight vector $\mathbf{w} = (w_1, ..., w_N)'$ to form a linear combination $\mathbf{w}'\mathbf{f}_t$ whose mean squared forecast error (MSFE) is minimized. One way to estimate $\mathbf{w}$ is to run the restricted least squares (RLS):

$$\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{2T}\sum_{t=1}^{T}\left(y_{t+1} - \mathbf{w}'\mathbf{f}_t\right)^2 \quad \text{subject to} \quad \mathbf{w}'\mathbf{1}_N = 1, \tag{1}$$

where $\mathbf{1}_N$ is an $N\times 1$ vector of ones. Alternatively, given the forecast error vector $\mathbf{e}_t = (e_{1t}, \ldots, e_{Nt})'$ with $e_{it} = y_{t+1} - f_{it}$ and its sample variance-covariance (VC) estimate: $\widehat{\mathbf{\Sigma}} \equiv T^{-1}\sum_{t=1}^{T}\mathbf{e}_t\mathbf{e}_t'$, we can follow Bates and Granger (1969) and consider the following minimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{2}\mathbf{w}'\widehat{\mathbf{\Sigma}}\mathbf{w} \quad \text{subject to} \quad \mathbf{w}'\mathbf{1}_N = 1. \tag{2}$$

Denote the solution to the above constrained optimization problem as $\widehat{\mathbf{w}}^{\text{BG}}$. When $\widehat{\mathbf{\Sigma}}$ is invertible, we can explicitly solve the problem to obtain the optimal solution:

$$\widehat{\mathbf{w}}^{\text{BG}} = \left(\mathbf{1}_N'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{1}_N\right)^{-1}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{1}_N. \tag{3}$$

The two formulations in (2) and (1) are numerically equivalent, as demonstrated by Granger and Ramanathan (1984). Apparently, the requirement of the invertibility of $\widehat{\mathbf{\Sigma}}$ is not innocuous in high dimensional setting, and in fact $\widehat{\mathbf{\Sigma}}$ is always singular if $N > T$.

Despite the MSFE-optimality of the above Bates and Granger's combined forecasts, the optimal weights often do not yield the best performed forecast in empirical researches when they are replaced with their sample estimates, and simple average (namely, *equal-weight* forecast combination) often performs better. This empirical fact has been best known as the "*forecast combination puzzle*", which was first noted by Clemen (1989) and formally named by Stock and Watson (2004). In particular, Clemen (1989) reviews the literature up to the late 1980s and notes that over a large number of papers averaging forecasts appear to be a more robust procedure than optimal combination in practice. A reasonable explanation suggests that the error on the estimation of the weights can be large and thus dominate the gains from the use of optimal combination; see, e.g., Smith and Wallis (2009) and Claeskens et al. (2016). A lesson learned from this literature is that it is not wise to include all possible variables and it is important to limit the number of parameters to be estimated so as to reduce the parameter estimation error. This has led to the adoption of various shrinkage and variable selection techniques in the literature. For example, Elliott et al. (2013) propose a complete subset regression (CSR) approach to forecast combinations by using equal weights to combine forecasts based on the same number of predictive variables. They show that in many cases subset regression combinations amount to a form of shrinkage that is more general than the conventional variable-by-variable shrinkage implied by ridge regression. In contrast, Diebold and Shin (2019) propose partially egalitarian Lasso (peLASSO) procedures that discard some forecasts and then select and shrink the remaining forecasts toward the equal weights. In essence, both Elliott et al. (2013) and Diebold and Shin (2019) assign two distinct weights to two subsets of models: zero weight to a large subset of the models and equal or roughly equal weights to the remaining subset of the models.

In this paper, we extend the ideas of Elliott et al. (2013) and Diebold and Shin (2019) and propose a new shrinkage technique for forecast combinations. Specifically, we propose to minimize the squared ℓ_2-norm of the weight vector subject to a relaxed version of the first order conditions from the minimization problem in (2), yielding the ℓ_2-relaxation problem. The strategy is similar in spirit to the ℓ_1-relaxation in Dantzig selector *a la* Candes and Tao (2007). Interestingly, the ℓ_2-relaxed optimal forecasts incorporate both the simple average (equal-weight) strategy and the conventional optimal weighting scheme as special cases by setting the tuning parameter to be sufficiently large or zero. A proper choice of the tuning parameter can achieve the usual bias and variance trade-off and deliver combined forecasts with roughly equal groupwise weights when the variance-covariance (VC) matrix of the individual forecasts exhibit a certain latent group structure. This is consistent with the intuition that one should assign the same weights to all individuals within the same group and potentially distinct weights to individuals in different groups when the forecast error VC matrix exactly exhibits a block equicorrelation structure. When the VC matrix is contaminated with a noisy component, we show that the resultant ℓ_2-relaxed optimal weights are close to the infeasible groupwise equal weights. We argue that the approximate latent group structure exist in many forecast combination problems. For example, it is present when the forecasting errors exhibit a factor structure as in Hsiao and Wan (2014) and Chan and Pauwels (2018), and the factor loadings exhibit certain latent group structure directly or are approximable by a fewer number of values. It is also present if one considers forecast combinations based on large number of forecast models (say, $2^{10} = 1024$) with a fixed number of predictive regressors (say, $p = 10$). In the latter case, we argue that the p regressors serve as a part of the "latent" factors.

It is worth mentioning that various shrinkage and machine learning techniques have been applied to the forecasting literature since the pioneer work of Tibshirani (1996). See, e.g., Li and Chen (2014), Conflitti et al. (2015), Konzen and Ziegelmann (2016), Stasinakis et al. (2016), Bayer (2018), Wilms et al. (2018), Kotchoni et al. (2019), Coulombe et al. (2020), Roccazzella et al. (2020). We complement these works by considering an ℓ_2-relaxation of the regularized weights estimation problem, which exhibit certain optimality properties when the forecast error VC matrix can be decomposed into the sum of a low-rank matrix plus a noisy matrix.

Our paper is also related to the recent literature on latent group structures in panel data analysis; see, e.g., Bonhomme and Manresa (2015), Su et al. (2016), Bonhomme et al. (2017), Su and Ju

(2018), Su et al. (2019), Vogt and Linton (2017), and Vogt and Linton (2020). Except Bonhomme and Manresa (2015) and Bonhomme et al. (2017), all these previous works focus on the recovery of the latent group structures. In this paper, we simply assume that the dominant term in the forecast error VC matrix exhibits a community or latent group structure. But we do not need to recover the membership for each individual forecast.

Lastly, our paper is related to the vast literature on portfolio optimization as well; see Ledoit and Wolf (2004), Disatnik and Katz (2012), Fan et al. (2012), Fan et al. (2016), Ledoit and Wolf (2017), and Ao et al. (2019), among many others. In particular, Ledoit and Wolf (2004) use Bayesian methods for shrinking the sample correlation matrix to an equicorrelated target and showed that this helps select portfolios with low volatility compared to those based on the sample correlation; and Ledoit and Wolf (2017) promote a nonlinear shrinkage estimator that is more flexible than previous linear shrinkage estimators and has just the right number of free parameters. Apparently, our method can also be used to estimate the optimal portfolios.

The rest of the paper is organized as follows. In Section 2, we introduce the ℓ_2-relaxation primal problem and its dual problem, and discuss their properties without imposing any structure on the VC matrix of the forecast errors. In Section 3, we consider the latent group structure in the VC matrix and study the finite sample (numerical) properties of the estimators in both the dual and primal problems. In Section 4, we study the asymptotic properties of the estimators in both the dual and primal problems and establish the asymptotic optimality of the ℓ_2-relaxed estimator of the combination weight. Section 5 reports some Monte Carlo simulation results. In Section 6, we apply the new method to two datasets. Section 7 concludes. We provide the proofs of all theoretical results in the appendix. Some additional technical results are contained in the online supplement.

Notation: For a random variable x, its population mean is $E[x]$. For a sample $(x_1, \ldots, x_T)$, its sample mean is $\mathbb{E}_T[x_t] = T^{-1}\sum_{t=1}^{T} x_t$. Plain b denotes a scalar, boldface lowercase $\mathbf{b}$ denotes a vector, and boldface uppercase $\mathbf{B}$ denotes a matrix. The ℓ_1-norm of $\mathbf{b}$ is denoted as $\|\mathbf{b}\|_1 = \sum_{i=1}^{n} |b_i|$ and the ℓ_2-norm as $\|\mathbf{b}\|_2 = \left(\sum_{i=1}^{n} b_i^2\right)^{1/2}$. For a generic $n \times m$ matrix $\mathbf{B}$, we denote $\mathbf{B}_{i\cdot}$ as the i-th row ($1 \times m$ vector), and $\mathbf{B}_{\cdot j}$ as the j-th column ($n \times 1$ vector). $\phi_{\max}(\mathbf{B})$ and $\phi_{\min}(\mathbf{B})$ are the largest and the smallest eigenvalues, respectively. The spectral norm $\|\mathbf{B}\|_{\mathrm{sp}} = \phi_{\max}^{1/2}(\mathbf{B}'\mathbf{B})$, the sup-norm $\|\mathbf{B}\|_\infty = \max_{i \le n, j \le m} |b_{ij}|$, and the maximum column ℓ_2 matrix norm $\|\mathbf{B}\|_{c2} = \max_{j \le m} \|\mathbf{B}_{\cdot j}\|_2$. $\mathbf{0}_n$ and $\mathbf{1}_n$ are $n \times 1$ vectors of zeros and ones, respectively, and $\mathbf{I}_n$ is an $n \times n$ identity matrix. "w.p.a.1" is short for "with probability approaching one" in asymptotic statements. We write $a \asymp b$ when both a/b and b/a are stochastically bounded. $a \wedge b = \min\{a, b\}$.

The cross-sectional units are indexed by $i \in [N] := \{1, \ldots, N\}$. For a generic index set $\mathcal{G} \subset [N]$, we denote $|\mathcal{G}|$ as the cardinality of $\mathcal{G}$, and $\mathbf{b}_{\mathcal{G}} = (b_i)_{i \in \mathcal{G}_k}$ as the $|\mathcal{G}|$-dimensional subvector. We call a vector of *similar sign* if no pair of its elements takes the opposite signs. Formally, an n-vector $\mathbf{b}$ is of similar sign if

$$\mathbf{b} \in \mathbb{S}^n := \{\mathbf{b} \in \mathbb{R}^n : b_i b_j \ge 0 \ \text{ for all } i, j \in [N]\}.$$

Let $\mathcal{G}_1, \ldots, \mathcal{G}_K$ be a partition of $[N]$, and denote $N_k = |\mathcal{G}_k|$. We call a generic N-vector $\mathbf{b}$ of *similar sign within all groups* if $\mathbf{b} \in \mathbb{S}^{\mathrm{all}} := \left\{\mathbf{b} \in \mathbb{R}^N : \mathbf{b}_{\mathcal{G}_k} \in \mathbb{S}^{N_k}, \text{ for all } k \in [K]\right\}$, and define

$\tilde{\mathbb{S}}^{\text{all}} := \left\{\mathbf{b} \in \mathbb{S}^{\text{all}} : \mathbf{1}_N'\mathbf{b} = 0\right\}$ with a further restriction that the vector $\mathbf{b}$ adds up as 0.

2 ℓ_2-relaxation for Optimal Forecast Combination

In this section, we introduce the idea of ℓ_2-relaxation to the classical forecast combination problem. We discuss the low-dimensional case first and then the high-dimensional case.

2.1 ℓ_2-relaxation in the low dimensional case

To solve the constrained optimization problem in (2), we rewrite it as the unconstrained Lagrangian problem:

$$\frac{1}{2}\mathbf{w}'\widehat{\mathbf{\Sigma}}\mathbf{w} + \gamma\left(\mathbf{w}'\mathbf{1}_N - 1\right), \tag{4}$$

where γ is the Lagrangian multiplier. If $\widehat{\mathbf{\Sigma}}$ is invertible, the explicit solution is given in (3). Nevertheless, $\widehat{\mathbf{\Sigma}}$ is frequently non-invertible in high-dimensional settings. In this case, (2) has multiple solutions, all of which satisfy the Kuhn-Karush-Tucker (KKT) conditions:

$$\begin{aligned} \widehat{\mathbf{\Sigma}}\mathbf{w} + \gamma\mathbf{1}_N &= \mathbf{0}_N \\ \mathbf{w}'\mathbf{1}_N - 1 &= 0. \end{aligned} \tag{5}$$

The non-unique solution due to the singularity of $\widehat{\mathbf{\Sigma}}$ is parallel to the familiar multi-collinearity issue in ordinary least squares (OLS) regression. To motivate ℓ_2-relaxation, we take a digression and consider the generic OLS regression in the following example.

Example 1. The OLS seeks a parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ that minimizes the sum of squared residuals (SSR) $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$, where $\mathbf{y}$ is a T-vector of dependent variable and $\mathbf{X}$ is a $T \times p$ matrix of independent variable. When columns in $\mathbf{X}$ are collinear, any solution that satisfies the first-order condition $\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = \mathbf{X}'\mathbf{y}$ minimizes the SSR. One solution is the well-known ridge regression (Tikhonov, 1977): $\widehat{\boldsymbol{\theta}}^{\text{ridge}} = \min_{\boldsymbol{\theta}\in\mathbb{R}^p}\left\{\frac{1}{2T}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2\right\}$, where λ is a tuning parameter. Alternatively, one can set up a criterion to compare the multiple solutions. For example, under the ℓ_2-norm, one of the most used criteria, the problem can be written as

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 \quad \text{subject to} \quad \mathbf{X}'\mathbf{X}\boldsymbol{\theta} = \mathbf{X}'\mathbf{y}. \tag{6}$$

Unlike the ridge regression, no tuning parameter is involved in (6). Apparently, when $\mathbf{X}'\mathbf{X}$ is nonsingular, the solution to the last minimization problem is given by the OLS solution $\widehat{\boldsymbol{\theta}}^{\text{OLS}}$ which is the only feasible point satisfying the constraint in (6).

Following the spirit of (6), a unique solution to (2) can be found by

$$\min_{(\mathbf{w},\gamma)\in\mathbb{R}^{N+1}} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{subject to} \quad \mathbf{w}'\mathbf{1}_N = 1 \text{ and } \widehat{\mathbf{\Sigma}}\mathbf{w} + \gamma\mathbf{1}_N = \mathbf{0}_N. \tag{7}$$

The above minimization problem does not require $\widehat{\boldsymbol{\Sigma}}$ being invertible. If $\widehat{\boldsymbol{\Sigma}}$ is indeed invertible, the unique solution to (7) is the same as (3).

2.2 ℓ_2-relaxation in the high dimensional case

Potential problems arise when N is large relatively to the time dimension T. "High dimensional" here means that the number of unknown parameters, in our context N, is comparable to the sample size T. For example, we may allow $N/T \to c \in (0, \infty]$ as $(N, T) \to \infty$.

Consider the moderate case where N is comparable to T but $N < T$, say $N = 80$ and $T = 100$. Even if $\widehat{\boldsymbol{\Sigma}}$ is non-singular, a few small sample eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ are likely to be close to zero, leading to a numerically unstable solution from (3). The numerical instability is due to the fact that the dimension of $\mathbf{w}$ also gets large in the case of large N, but the singleton feasible point is solely decided by $\widehat{\boldsymbol{\Sigma}}$ in the N-equation linear system $\widehat{\boldsymbol{\Sigma}}\mathbf{w} + \gamma\mathbf{1}_N = \mathbf{0}_N$.

An idea to stabilize the numerical solution is to expand the feasible set. Inspired by the Dantzig selector of Candes and Tao (2007) and the relaxed empirical likelihood of Shi (2016), we consider relaxing the sup-norm of the KKT condition as follows:

$$\min_{(\mathbf{w},\gamma)\in\mathbb{R}^{N+1}} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{subject to} \;\; \mathbf{w}'\mathbf{1}_N = 1 \text{ and } \|\widehat{\boldsymbol{\Sigma}}\mathbf{w} + \gamma\mathbf{1}_N\|_\infty \leq \tau, \tag{8}$$

where τ is a tuning parameter to be specified by the user. We call the problem in (8) the ℓ_2-*relaxation primal problem*, and denote the solution to (8) as $\widehat{\mathbf{w}} = \widehat{\mathbf{w}}_\tau$, where we frequently suppress the dependence of $\widehat{\mathbf{w}}$ on τ for conciseness.

ℓ_2-relaxation includes the low dimensional problem (7) as a special case when τ is set as zero. When $\tau > 0$, we relax the high dimensional component of the KKT condition (5). Since (2) always has a solution, constraints in (8) are feasible for any $\tau \geq 0$. Moreover, the solution to (8) is unique because the objective is a strictly convex function and the feasible set is a closed convex set. Therefore, while (2) may have multiple solutions, the objective function in (8) selects in the feasible set the solution with the smallest ℓ_2-norm.

To proceed, we consider the dual problem of (8). The result is stated in the following lemma.

Lemma 1. *Let the primal problem be given in (8). Then the dual problem is*

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \left\{ \frac{1}{2}\boldsymbol{\alpha}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\alpha} + \frac{1}{N}\mathbf{1}_N'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} + \tau\|\boldsymbol{\alpha}\|_1 - \frac{1}{2N} \right\} \quad \textit{subject to} \;\; \mathbf{1}_N'\boldsymbol{\alpha} = 0, \tag{9}$$

where $\widehat{\mathbf{A}} = \left(\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'\right)\widehat{\boldsymbol{\Sigma}}$ *is the demeaned version of* $\widehat{\boldsymbol{\Sigma}}$. *Denote* $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_\tau$ *as a solution to the dual problem in (9). Then the solution of* $\mathbf{w}$ *to the primal problem in (8) is given by*

$$\widehat{\mathbf{w}} = \widehat{\mathbf{w}}_\tau = \widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}} + \frac{\mathbf{1}_N}{N}. \tag{10}$$

Remark 2.1. When $\tau = 0$ in (9), it is the dual problem of (7). When $\tau > 0$, it is a constrained ℓ_1-penalized optimization where the criterion function is a summation of a quadratic form of $\boldsymbol{\alpha}$, a linear combination of $\boldsymbol{\alpha}$, and the ℓ_1-norm of $\boldsymbol{\alpha}$, while the constraint is linear in $\boldsymbol{\alpha}$. The dual problem is instrumental in our theoretical analyses due to its similarity to the familiar Lasso, the ℓ_1-penalized sparse regression problem (Tibshirani, 1996).

Remark 2.2. (Zhentao to Liangjun: you wrote "Noting that $\widehat{\mathbf{A}}$ is singular no matter whether $\widehat{\boldsymbol{\Sigma}}$ is or not, the solution to the dual problem in (9) is not unique in general." Zhentao thinks that if $\widehat{\boldsymbol{\Sigma}}$ is of full rank, because of the extra restriction $\mathbf{1}_N'\boldsymbol{\alpha} = 0$, the solution $\widehat{\boldsymbol{\alpha}}$ is still unique. When the VC is $\boldsymbol{\Sigma}^*$, whose rank is much lower than N, then $\widehat{\boldsymbol{\alpha}}$ is non-unique.) Since rank$\left(\widehat{\mathbf{A}}\right) \leq$ rank$(\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N') = N-1$, the singularity may induce multiple solutions to the dual (9). Despite this, the uniqueness of $\widehat{\mathbf{w}}$ as a solution to the primal problem in (8) implies $\widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}}^{(1)} = \widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}}^{(2)}$ for any $\widehat{\boldsymbol{\alpha}}^{(1)}$ and $\widehat{\boldsymbol{\alpha}}^{(2)}$ that solve (9). So in practice, it is sufficient to find any solution to the dual problem to recover the same $\widehat{\mathbf{w}}$ in the primal.

Remark 2.3. The tuning parameter τ plays a crucial role for the ℓ_2-relaxation problem. Interestingly, the penalized scheme in (8) incorporates the two extremes, simple average and optimal weighting, as special cases.

(a) If $\tau = 0$, the solution $\widehat{\mathbf{w}}$ is characterized by the KKT conditions in (5); if, in addition, $\widehat{\boldsymbol{\Sigma}}$ is invertible, the unique solution is then given by the optimal weighting $\widehat{\mathbf{w}}^{\mathrm{BG}}$ defined in (3).

(b) If τ is sufficiently large, say $\tau \geq \max_{i\in[N]}\left|\frac{1}{N}\widehat{\boldsymbol{\Sigma}}_{i\cdot}'\mathbf{1}_N\right|$, then $\frac{1}{2}\boldsymbol{\alpha}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\alpha} + \frac{1}{N}\mathbf{1}_N'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} + \tau\left\|\boldsymbol{\alpha}\right\|_1 \geq 0$ and the equality holds if and only if $\boldsymbol{\alpha} = \mathbf{0}_N$. Since $\boldsymbol{\alpha} = \mathbf{0}_N$ is feasible (i.e., satisfying the constraints) for the dual problem, the solution to the dual problem must be given by $\widehat{\boldsymbol{\alpha}} = \mathbf{0}_N$. Then $\widehat{\mathbf{w}} = N^{-1}\mathbf{1}_N$ by (10). That is, the simple average is optimal for the primal problem in (8) provided τ is sufficiently large.

(c) The optimal weighting strategy often does not perform well in practice because the estimation of the optimal weights often cause biases and bring in large estimation errors. If the tuning parameter τ is chosen properly, the ℓ_2-relaxation can achieve the right balance between the optimal weighting and the simple average by exploring their bias-variance trade-off. Figure 1 demonstrates the bias and variance trade-off under a range of τ, where the DGP is described in Section B.1. As can be seen clearly from the figure, the ℓ_2-relaxed weights associated with a small value of τ tends to have no or small biases but large variances whereas those associated with a large value of τ tends to have big biases but small variances. In the middle, there is a large range of values for τ where the combined forecast yields MSFE that is smaller than either that for the simple average estimator (attainable for sufficiently large τ) or the Bates-Granger optimally combined estimator (attainable for $\tau = 0$).

In the next section, we assume that $\widehat{\boldsymbol{\Sigma}}$ has a certain structure and then examine its implications on the optimal forecast combination based on the ℓ_2-relaxation.

Figure 1: Bias-variance Trade-off under different τ values

3 Latent Group Structure and Its Implications

In this section, we impose some structure on $\widehat{\mathbf{\Sigma}}$ (or its population version) and then study its implications on the ℓ_2-relaxed estimates of the optimal weights.

3.1 Decomposition of $\widehat{\mathbf{\Sigma}}$ and latent group structure

Statistical analysis of high-dimensional problems often impose structures on the data generating process. For example, various variable selection methods such as Lasso (Tibshirani, 1996) and SCAD (Fan and Li, 2001) are motivated from regressions with sparsity, meaning most of the regression coefficients are either exactly zero or close to zero. Similarly, in large covariance estimation various structures have been considered in the literature. For example, Bickel and Levina (2008) assume many off-diagonal elements to be zero; Engle and Kelly (2012) assume a block equicorrelation structure. In portfolio optimization, Ledoit and Wolf (2004) use Bayesian methods for shrinking the sample correlation matrix to an equicorrelated target.

The burgeoning literature of latent group structures in panel data analyses is an alternative way to reduce dimensions. While the optimization problem (8) is formulated for a generic covariance matrix $\widehat{\mathbf{\Sigma}}$, to analyze it in depth in high dimension we assume $\widehat{\mathbf{\Sigma}} = \{\widehat{\Sigma}_{ij}\}_{i,j\in[N]}$ can be approximated by a block equicorrelation matrix:[1]

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^* + \widehat{\mathbf{\Sigma}}^e, \tag{11}$$

where $\widehat{\mathbf{\Sigma}}^* = \{\widehat{\Sigma}^*_{ij}\}_{i,j\in[N]}$ is a block equicorrelation matrix and $\widehat{\mathbf{\Sigma}}^e = \{\widehat{\Sigma}^e_{ij}\}_{i,j\in[N]}$ denotes the devia-

[1] Alternatively, we can also assume that the population version of $\widehat{\mathbf{\Sigma}}^*$ exhibits a block equicorrelation structure. In this paper, we assume $\widehat{\mathbf{\Sigma}}^*$ itself has such a structure and the difference between the two can be captured by the deviation matrix $\widehat{\mathbf{\Sigma}}^e$.

tion of $\widehat{\mathbf{\Sigma}}$ from the block equicorrelation matrix. We write

$$\widehat{\mathbf{\Sigma}}^{*} = \mathbf{Z}\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{Z}' \tag{12}$$

where $\mathbf{Z} = \{Z_{ik}\}$ denotes an $N \times K$ binary matrix providing the cluster membership of each individual forecast, i.e., $Z_{ik} = 1$ if forecast i belongs to group $\mathcal{G}_k \subset [N]$ and $Z_{ik} = 0$ otherwise, and $\widehat{\mathbf{\Sigma}}^{\text{co}} = \{\widehat{\Sigma}^{\text{co}}_{kl}\}_{k,l\in[K]}$ is a $K \times K$ symmetric positive definite matrix. Here, the superscript "co" stands for "core". Note that

$$\widehat{\Sigma}^{*}_{ij} = \widehat{\Sigma}^{\text{co}}_{kl} \text{ if } i \in \mathcal{G}_k \text{ and } j \in \mathcal{G}_l.$$

An econometrician observes $\widehat{\mathbf{\Sigma}}$ from the data but not $\widehat{\mathbf{\Sigma}}^{*}$. We will be precise about the definition of "approximation" for $\widehat{\mathbf{\Sigma}}^{e}$ in Assumption 1 in the next section. Remind $N_k := |\mathcal{G}_k|$ is the number of individuals in the k-th group, and $\sum_{k=1}^{K} N_k = N$.

Our theory below does not require the exact knowledge on the membership matrix $\mathbf{Z}$ that contains the membership information. For ease of notation and perhaps after the reordering of the N forecast units, we assume that

$$\widehat{\mathbf{\Sigma}}^{*} = (\widehat{\Sigma}^{\text{co}}_{kl} \cdot \mathbf{1}_{N_k}\mathbf{1}'_{N_l})_{k,l\in[K]}, \tag{13}$$

in which the units in the same group cluster together. The theory to be developed is irrelevant to the ordering of the forecast units. The re-ordering is for the convenience of notation only.

We now motivate the above decomposition through two examples.

Example 2. In a recent paper, Chan and Pauwels (2018) assume the existence of a "best" unbiased forecast f_{0t} of variable y_{t+1} with associated forecast error e_{0t}, and the forecast error e_{it} of model i can be decomposed as

$$e_{it} = e_{0t} + u_{it},$$

where e_{0t} represents the forecast error from the best forecasting model, and u_{it} represents the difference in forecast error between the best forecasting model and the model i. Assuming that $E\left[u_{it}\right] = 0$ and $E\left[e_{0t}u_{it}\right] = 0$ for each i, the VC of $\mathbf{e}_t$ can be written as $\mathbf{\Sigma} := E\left[\mathbf{e}_t\mathbf{e}'_t\right] = E\left[e^2_{0t}\right]\mathbf{1}_N\mathbf{1}'_N + E[\mathbf{u}_t\mathbf{u}'_t]$, where $\mathbf{u}_t = (u_{1t}, ..., u_{Nt})'$. At the sample level, we have $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^{*} + \widehat{\mathbf{\Sigma}}^{e}$, where

$$\widehat{\mathbf{\Sigma}} = \mathbb{E}_T\left[\mathbf{e}_t\mathbf{e}'_t\right], \ \widehat{\mathbf{\Sigma}}^{*} = \mathbb{E}_T\left[e^2_{0t}\right]\mathbf{1}_N\mathbf{1}'_N, \text{ and } \widehat{\mathbf{\Sigma}}^{e} = \mathbb{E}_T[\mathbf{u}_t\mathbf{u}'_t] + \mathbf{1}_N\mathbb{E}_T[e_{0t}\mathbf{u}'_t] + \mathbb{E}_T[e_{0t}\mathbf{u}_t]\mathbf{1}'_N.$$

In this case, all N forecast units belong to the same group $\mathcal{G}_1$ as $\text{rank}(\widehat{\mathbf{\Sigma}}^{*}) = 1$.

Example 3. Consider each individual forecast f_{it} is generated from a factor model

$$f_{it} = \boldsymbol{\lambda}'_{g_i}\boldsymbol{\eta}_t + u_{it}, \tag{14}$$

where $\boldsymbol{\lambda}_{g_i}$ is a $q \times 1$ vector of factor loadings, $\boldsymbol{\eta}_t$ is a $q \times 1$ vector of latent factors, and u_{it} is the idiosyncratic shock. Here g_i denotes individual i's membership, i.e., it takes value k if individual i

belongs to group $\mathcal{G}_k$ for $k \in [K]$ and $i \in [N]$, where K is usually much smaller than N. Similarly, assume $y_{t+1} = \boldsymbol{\lambda}_y' \boldsymbol{\eta}_t + u_{y,t+1}$. Assume that $E\left[u_{it} | \boldsymbol{\eta}_t\right] = 0$ and $E\left[u_{y,t+1} | \boldsymbol{\eta}_t\right] = 0$ and $E\left[u_{it} u_{y,t+1} | \boldsymbol{\eta}_t\right] = 0$.[2] Let $\boldsymbol{\eta}_t^{\dagger} = (\boldsymbol{\eta}_t', u_{y,t+1})'$, $\boldsymbol{\lambda}_{g_i}^{\dagger} = ((\boldsymbol{\lambda}_y - \boldsymbol{\lambda}_{g_i})', 1)'$, $\mathbf{e}_t = (e_{1t}, ..., e_{Nt})'$. For simplicity, we also assume conditional homoskedasticity $\operatorname{var}(\mathbf{u}_t | \boldsymbol{\eta}_t) = \boldsymbol{\Omega}_u$ and the factor loadings are nonstochastic. Then individual i's forecast error is

$$e_{i,t} := y_{t+1} - f_{i,t} = \left[(\boldsymbol{\lambda}_y - \boldsymbol{\lambda}_{g_i})' \boldsymbol{\eta}_t + u_{y,t+1}\right] - u_{i,t} = \lambda_{g_i}^{\dagger\prime} \boldsymbol{\eta}_t^{\dagger} - u_{i,t}$$

or in vector form: $\mathbf{e}_t = \boldsymbol{\Lambda}^{\dagger} \boldsymbol{\eta}_t^{\dagger} - \mathbf{u}_t$, where $\boldsymbol{\Lambda}^{\dagger} \colon = \left(\lambda_{g_1}^{\dagger}, \ldots, \lambda_{g_N}^{\dagger}\right)'$. Then

$$\boldsymbol{\Sigma} := E\left[\mathbf{e}_t \mathbf{e}_t'\right] = \boldsymbol{\Lambda}^{\dagger} E[\boldsymbol{\eta}_t^{\dagger} \boldsymbol{\eta}_t^{\dagger\prime}] \boldsymbol{\Lambda}^{\dagger\prime} + \boldsymbol{\Omega}_u.$$

Decompose the sample VC as $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}^{*} + \widehat{\boldsymbol{\Sigma}}^{e}$, where

$$\widehat{\boldsymbol{\Sigma}} = \mathbb{E}_T\left[\mathbf{e}_t \mathbf{e}_t'\right], \ \widehat{\boldsymbol{\Sigma}}^{*} = \boldsymbol{\Lambda}^{\dagger} \mathbb{E}_T[\boldsymbol{\eta}_t^{\dagger} \boldsymbol{\eta}_t^{\dagger\prime}] \boldsymbol{\Lambda}^{\dagger\prime}, \text{ and } \widehat{\boldsymbol{\Sigma}}^{*} = \mathbb{E}_T[\mathbf{u}_t \mathbf{u}_t'] - \boldsymbol{\Lambda}^{\dagger} \mathbb{E}_T[\boldsymbol{\eta}_t^{\dagger} \mathbf{u}_t'] - \mathbb{E}_T[\mathbf{u}_t \boldsymbol{\eta}_t^{\dagger\prime}] \boldsymbol{\Lambda}^{\dagger\prime}.$$

In this case, $\operatorname{rank}(\widehat{\boldsymbol{\Sigma}}^{*}) = (q+1) \wedge K$, and by construction $\widehat{\Sigma}_{kl}^{\text{co}} = \lambda_k^{\dagger\prime} \mathbb{E}_T[\boldsymbol{\eta}_t^{\dagger} \boldsymbol{\eta}_t^{\dagger\prime}] \lambda_l^{\dagger}$ for $k, l \in [K]$, and $\widehat{\Sigma}_{ij}^{*} = \widehat{\Sigma}_{kl}^{\text{co}}$ if $i \in \mathcal{G}_k$ and $j \in \mathcal{G}_l$.

Remark 3.1. We emphasize that our theory below does not require the knowledge on the group membership for each individual forecasts. If one believes in the multi-factor structure in (14), one can always conduct the principal component analysis (PCA) to estimate the factor loadings and factors first. Then we can apply either the K-means algorithm or the sequential binary segmentation algorithm of Wang and Su (2020) to the estimates of the factor loadings to estimate the group membership. Alternatively, we can consider the regression of f_{it} on the estimated factors and apply the C-Lasso of Su et al. (2016) or other methods to recover the group membership. The advantage of our ℓ_2-relaxation is that it bypasses the factor structure or the group membership recovery and allows us to directly work with the sample moments.

Remark 3.2. It is worth mentioning, Hsiao and Wan (2014) also assume that the forecast errors exhibit a multi-factor structure. But they do not assume the presence of K latent groups in the N factor loadings and write $\boldsymbol{\lambda}_i$ in place of $\boldsymbol{\lambda}_{g_i}$ in Example 3. In the absence of the latent group structure among the factor loadings $\{\boldsymbol{\lambda}_i\}_{i \in [N]}$, we can follow the lead of Bonhomme et al. (2017) and consider their discretization. For simplicity, we assume that $\boldsymbol{\lambda}_i$ lies in a compact parameter space Λ in $\mathbb{R}^r$. Then for each $i \in [N]$, there exists K, $g_i \in [K]$ and $\boldsymbol{\lambda}_{g_i} \in \Lambda$ such that $\|\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_{g_i}\|_2 \leq \delta_K$. As K diverges to infinity, the approximation error δ_K can be made as small as possible. As a result, we can continue to decompose the i-th forecast error in Example 3 as $e_{i,t} = \boldsymbol{\lambda}_{g_i}^{\dagger\prime} \boldsymbol{\eta}_t^{\dagger} - u_{i,t}^{a}$, where $\boldsymbol{\lambda}_{g_i}^{\dagger} = ((\boldsymbol{\lambda}_y - \boldsymbol{\lambda}_{g_i})', 1)'$ and $u_{i,t}^{a}$ now contains the discretization error.

[2] Other than those s factors in $\boldsymbol{\eta}_t$, the additional latent factor $u_{y,t+1}$ in y_{t+1} is unforeseeable at time t. In other words, given the information set $\mathcal{I}_t$ that contains $\left(\{f_{it}\}_{i \in [N]}, \boldsymbol{\eta}_t\right)$ and $\mathbf{u}_t$ at time t, the error $u_{y,t+1} = y_{t+1} - \boldsymbol{\lambda}_y' \boldsymbol{\eta}_t = y_{t+1} - E\left(y_{t+1} | \mathcal{I}_t\right)$ must be orthogonal to $\mathcal{I}_t$. Then $E\left[u_{it} u_{y,t+1} | \mathcal{I}_t\right] = 0$ implies $E\left[u_{it} u_{y,t+1} | \boldsymbol{\eta}_t\right] = 0$ by the law of iterated expectations.

Remark 3.3. Even though we motivate the latent group structure via the use of approximate factor models in the above two examples, one should not have the impression that our method is only working for factor models. In fact, many forecast errors have an implicit multi-factor structure. See the next example.

Example 4. Suppose that the outcome variable y_{t+1} is generated via the process: $y_{t+1} = \mathbf{x}_t'\boldsymbol{\beta}^0 + u_{t+1}$, where $t = -(T_0-1), ..., -1, 0, 1, ...$, $\mathbf{x}_t = (x_{1,t}, ..., x_{p,t})'$ is a $p \times 1$ vector of potential predictive variables, $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_p^0)'$ is $p{\times}1$ vector of regression coefficients that is s-sparse, namely, $\left|\boldsymbol{\beta}^0\right|_0 = s$. Without loss of generality, we can assume that $\beta_1^0, ..., \beta_s^0$ are nonzero and $\beta_{s+1}^0, ..., \beta_p^0$ are all zeros. Let $S^0 = [s]$ denote the set of regressors whose true regression coefficients are nonzero. Let $S_i \subset [p]$ index the set of regressors included in forecasting model i for $i \in [N]$. Let $\mathbf{y}_t = (y_{-(T_0-1)}, ..., y_t)', \mathbf{u}_t = (u_{-(T_0-1)}, ..., u_t)'$, and $\mathbf{X}_{t-1}$= $(\mathbf{x}_{-T_0}, ..., \mathbf{x}_{t-1})'$ for $t = 0, 1, ..., T$. Let $\mathbf{X}_{S_i,t}$ denote a $(T_0+t) \times |S_i|$ matrix that includes regressors specified in the index set S_i. Note that $\mathbf{y}_{t+1} = \mathbf{X}_{t-1}\boldsymbol{\beta}^0 + u_{t+1}$. Let $\widehat{\boldsymbol{\beta}}_{S_i,t} = (\mathbf{X}_{S_i,t}'\mathbf{X}_{S_i,t})^{-1}\mathbf{X}_{S_i,t}'\mathbf{y}_t$, and $\widehat{\boldsymbol{\beta}}_{i,t}$ be a $p \times 1$ vector whose j-th element is given by the corresponding element in $\widehat{\boldsymbol{\beta}}_{S_i,t}$ if $j \in S_i$ and zero otherwise. Note that $\widehat{\boldsymbol{\beta}}_{i,t} = \mathcal{S}_i\widehat{\boldsymbol{\beta}}_{S_i,t}$, where $\mathcal{S}_i = \{\mathcal{S}_{i,jl}\}$ is a $p \times |S_i|$ matrix such that $\mathcal{S}_{i,jl} = 1$ if and only if $j \in S_i$ and the j-th regressor lies in the l-th position in $\mathbf{X}_{S_i,t}$ and $\mathcal{S}_{i,jl} = 0$ otherwise.

We consider two forecasting schemes: fixed window and rolling window.

(i) In the case of fixed window, the ith forecast of y_{t+1} is given by $\hat{f}_{it} \equiv \mathbf{x}_t'\widehat{\boldsymbol{\beta}}_{i,0}$ for $t = 1, ..., T$, with the associated forecast error given by $e_{it} = y_{t+1} - \mathbf{x}_t'\widehat{\boldsymbol{\beta}}_{i,0}$. Note that

$$e_{it} = y_{t+1} - \mathbf{x}_t'\widehat{\boldsymbol{\beta}}_{i,0} = u_{t+1} + \mathbf{x}_t'(\boldsymbol{\beta}^0 - \widehat{\boldsymbol{\beta}}_{i,0}).$$

Apparently, this is an exact $(p+1)$-factor model with factors $(\mathbf{x}_t', u_{t+1})'$ and factor loadings $((\boldsymbol{\beta}^0 - \widehat{\boldsymbol{\beta}}_{i,0})', 1)'$.

(ii) In the case of rolling window, we assume that $\widehat{\mathbf{Q}}_{S_iS_i,t} := \frac{1}{T_0+t}\mathbf{X}_{S_i,t}'\mathbf{X}_{S_i,t} \xrightarrow{p} \widehat{\mathbf{Q}}_{S_iS_i}$ and $\widehat{\mathbf{Q}}_{S_i,t} := \frac{1}{T_0+t}\mathbf{X}_{S_i,t}'\mathbf{X}_{t-1} \xrightarrow{p} \mathbf{Q}_{S_i}$ uniformly in (i,t) under some regularity conditions that include the covariance stationarity,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{S_i,t} &= (\mathbf{X}_{S_i,t}'\mathbf{X}_{S_i,t})^{-1}\mathbf{X}_{S_i,t}'(\mathbf{X}_{t-1}\boldsymbol{\beta}^0 + \mathbf{u}_t) \\
&= \widehat{\mathbf{Q}}_{S_iS_i,t}^{-1}\widehat{\mathbf{Q}}_{S_i,t}\boldsymbol{\beta}^0 + \widehat{\mathbf{Q}}_{S_iS_i,t}^{-1}\frac{1}{T_0+t}\mathbf{X}_{S_i,t}'\mathbf{u}_t \\
&= \mathbf{Q}_{S_iS_i}^{-1}\mathbf{Q}_{S_i}\boldsymbol{\beta}^0 + \zeta_{it}
\end{aligned}$$

where $\zeta_{it} = \widehat{\mathbf{Q}}_{S_iS_i,t}^{-1}\frac{1}{T_0+t}\mathbf{X}_{S_i,t}'\mathbf{u}_t + (\widehat{\mathbf{Q}}_{S_iS_i,t}^{-1}\widehat{\mathbf{Q}}_{S_i,t} - \mathbf{Q}_{S_iS_i}^{-1}\mathbf{Q}_{S_i})\boldsymbol{\beta}^0$. Then the forecast error is

$$\begin{aligned}
e_{it} &= y_{t+1} - \mathbf{x}_t'\widehat{\boldsymbol{\beta}}_{i,t} = y_{t+1} - \mathbf{x}_t'\mathcal{S}_i\widehat{\boldsymbol{\beta}}_{S_i,t} \\
&= u_{t+1} + \mathbf{x}_t'(\mathbf{I}_p - \mathcal{S}_i\mathbf{Q}_{S_iS_i}^{-1}\mathbf{Q}_{S_i})\boldsymbol{\beta}^0 + \epsilon_{it},
\end{aligned}$$

where $\epsilon_{it} = -\mathbf{x}_t'\mathcal{S}_i\zeta_{it}$. Therefore, we have an approximate $(p+1)$-factor model with factors $(\mathbf{x}_t', u_{t+1})'$ and factor loadings $(\boldsymbol{\beta}^{0\prime}(\mathbf{I}_p - \mathcal{S}_i\mathbf{Q}_{S_iS_i}^{-1}\mathbf{Q}_{S_i})', 1)'$.

In either case, imagine one has $p = 10$ regressors and $N = 2^{10} = 1024$ forecasting models under consideration. It is reasonable to model the forecast errors to have a multi-factor structure as above.

3.2 The oracle problem

Given the latent group structure, an oracle version of (2) is

$$\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{2}\mathbf{w}'\widehat{\boldsymbol{\Sigma}}^*\mathbf{w} \quad \text{subject to } \; \mathbf{w}'\mathbf{1}_N = 1, \tag{15}$$

where the infeasible block equicorrelation covariance matrix $\widehat{\boldsymbol{\Sigma}}^*$ replaces the sample covariance $\widehat{\boldsymbol{\Sigma}}$ in (2). In the oracle problem, the group structure can be identified by inspecting the values of the elements in $\widehat{\boldsymbol{\Sigma}}^*$. The rank of $\widehat{\boldsymbol{\Sigma}}^*$ is at most K due to the latent group pattern, leading to multiple and in fact an infinite number of solutions. Despite this, each solution yields the same optimal value for the primal objective function. Therefore it suffices to identify one solution. The counterpart of (7) here is

$$\min_{(\mathbf{w},\gamma)\in\mathbb{R}^{N+1}} \frac{1}{2}\left\|\mathbf{w}\right\|_2^2 \quad \text{subject to } \; \mathbf{w}'\mathbf{1}_N = 1 \text{ and } \widehat{\boldsymbol{\Sigma}}^*\mathbf{w} + \gamma\mathbf{1}_N = \mathbf{0}_N. \tag{16}$$

The ℓ_2-norm objective function produces the *within-group equally weighted solution*, which we write as

$$\underset{N\times 1}{\mathbf{w}_0^*} = \left(\frac{b_{01}^*}{N}\cdot\mathbf{1}_{N_1}', \cdots, \frac{b_{0K}^*}{N}\cdot\mathbf{1}_{N_K}'\right)', \tag{17}$$

for some $\mathbf{b}_0^* = (b_{01}^*, \ldots, b_{0K}^*)'$ satisfying $\sum_{k=1}^K N_k b_{0k}^* = N$. Let $r_k := N_k/N$ be the fraction of the k-th group members on the cross section. Denote $\mathbf{w}_{\mathbf{b}_0^*} = (b_{01}^* r_1, \ldots, b_{0K}^* r_K)'$ as the compressed version of $\mathbf{w}_0^*$. It is easy to verify that $\mathbf{w}_{\mathbf{b}_0^*} = \frac{(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K}{\mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K}$ with $\mathbf{1}_K'\mathbf{w}_{\mathbf{b}_0^*} = 1$, and

$$b_{0k}^* = \frac{[(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K]_k}{r_k\cdot\mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K}, \tag{18}$$

That is, b_{0k}^* solely depends on $(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}$ and r_k, as one might expect.

The oracle counterpart of the ℓ_2-relaxation primal problem in (8) is

$$\min_{(\mathbf{w},\gamma)\in\mathbb{R}^{N+1}} \frac{1}{2}\left\|\mathbf{w}\right\|_2^2 \quad \text{subject to } \; \mathbf{w}'\mathbf{1}_N = 1, \text{ and } \|\widehat{\boldsymbol{\Sigma}}^*\mathbf{w} + \gamma\|_\infty \leq \tau, \tag{19}$$

and its dual problem is

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^N} \left\{\frac{1}{2}\boldsymbol{\alpha}'\widehat{\mathbf{A}}^{*\prime}\widehat{\mathbf{A}}^*\boldsymbol{\alpha} + \frac{1}{N}\mathbf{1}_N'\widehat{\boldsymbol{\Sigma}}^*\boldsymbol{\alpha} + \tau\left\|\boldsymbol{\alpha}\right\|_1 - \frac{1}{2N}\right\} \quad \text{subject to } \; \mathbf{1}_N'\boldsymbol{\alpha} = 0, \tag{20}$$

where $\widehat{\mathbf{A}}^* = \left(\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'\right)\widehat{\boldsymbol{\Sigma}}^*$.

Remark 3.4. In general, $\widehat{\mathbf{A}}^*$ is not a symmetric matrix, though $\widehat{\boldsymbol{\Sigma}}^*$ is symmetric. The group structure in $\widehat{\boldsymbol{\Sigma}}^*$ implies that $\widehat{\mathbf{A}}_{i\cdot}^* = \widehat{\mathbf{A}}_{j\cdot}^*$ for all $i, j \in \mathcal{G}_k$ for some $k \in [K]$. Thus $\widehat{\mathbf{A}}^*\boldsymbol{\alpha}^{(1)} = \widehat{\mathbf{A}}^*\boldsymbol{\alpha}^{(2)}$

for any $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ such that $\boldsymbol{\alpha}^{(1)}_{\mathcal{G}_k} - \boldsymbol{\alpha}^{(2)}_{\mathcal{G}_k} = \mathbf{0}_{N_k}$ for all $k \in [K]$. For any $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}^*_\tau$ (the subscript τ stresses its dependence on τ) that solves (20), Lemma 1 implies that the *oracle optimal weighting* is given by

$$\mathbf{w}^*_\tau = \widehat{\mathbf{A}}^* \boldsymbol{\alpha}^*_\tau + \frac{\mathbf{1}_N}{N}. \tag{21}$$

The restricted quadratic optimization problem (19) produces a unique solution $\mathbf{w}^*_\tau$ whose within-group weights must be equal. It is obvious that $\mathbf{w}^*_0$ is the special case of $\mathbf{w}^*_\tau$ when $\tau = 0$. As above, the solution to (20) is not unique. For any $\boldsymbol{\alpha}^{*(1)}_\tau$ and $\boldsymbol{\alpha}^{*(2)}_\tau$ that solve (20), we must have $\widehat{\mathbf{A}}^* \boldsymbol{\alpha}^{*(1)}_\tau = \widehat{\mathbf{A}}^* \boldsymbol{\alpha}^{*(2)}_\tau$ and $\boldsymbol{\alpha}^{*(1)}_{\mathcal{G}_k} - \boldsymbol{\alpha}^{*(2)}_{\mathcal{G}_k} = \mathbf{0}_{N_k}$. The penalization of the ℓ_1-norm precludes opposite signs within a group. We thus also have $\|\boldsymbol{\alpha}^{*(1)}_\tau\|_1 = \|\boldsymbol{\alpha}^{*(2)}_\tau\|_1$.

3.3 Numerical properties

In this section, we derive useful finite sample numerical properties which holds for any finite N. We first characterize $\boldsymbol{\alpha}^*_0$ that solves (21) with $\tau = 0$: $\mathbf{w}^*_0 = \widehat{\mathbf{A}}^* \boldsymbol{\alpha}^*_0 + \frac{\mathbf{1}_N}{N}$. The behavior of the oracle estimator is governed not only by the "core" covariance $\widehat{\boldsymbol{\Sigma}}^{\text{co}}$ but also by the relative size of each group r_k. Define a $K \times K$ matrix:

$$\widehat{\mathbf{A}}^{\text{co}} = \mathbf{R}^{1/2} \left(\mathbf{I}_K - \mathbf{1}_K \cdot \mathbf{r}' \right) \widehat{\boldsymbol{\Sigma}}^{\text{co}},$$

where $\mathbf{r} = (r_k)_{k=1}^K$ is a $K \times 1$ vector, $\mathbf{R} = \text{diag}(\mathbf{r})$ is the $K \times K$ diagonal matrix with elements of $\mathbf{r}$ on the diagonal. $\widehat{\mathbf{A}}^{\text{co}}$ is the *weighted demeaned core* of $\widehat{\mathbf{A}}^*$, just like $\widehat{\boldsymbol{\Sigma}}^{\text{co}}$ is the core of $\widehat{\boldsymbol{\Sigma}}^*$. Here "demeaning" is weighted by the relative size of the groups, in contrast to the plain demeaned $\left(\mathbf{I}_K - K^{-1}\mathbf{1}_K \mathbf{1}'_K\right) \widehat{\boldsymbol{\Sigma}}^{\text{co}}$. We construct $\widehat{\mathbf{A}}^{\text{co}}$ in this way so that the $K \times K$ core $\widehat{\mathbf{A}}^{\text{co}}$ and the corresponding $N \times N$ full-scale matrix $\widehat{\mathbf{A}}^*$ are connected as in Lemma 2(a) below. Define a K-vector $\mathbf{a} = (a_1, \ldots, a_K)'$, where each element is the within-group summation of $\{\alpha_i\}_{i\in[N]}$: $a_k = \sum_{i \in \mathcal{G}_k} \alpha_i$. Let $\mathbf{a}^*_0 = (a^*_{01}, \ldots, a^*_{0K})'$, where $a^*_{0k} = \sum_{i\in\mathcal{G}_k} \alpha^*_{0,i}$.

Define a $(K+1)\times K$ matrix $\tilde{\mathbf{A}}^{\text{co}} = \left(\widehat{\mathbf{A}}^{\text{co}\prime} \quad \mathbf{1}_K \right)'$ and a $(K+1)$-vector $\tilde{\mathbf{b}}^{\text{co}} = \left((\mathbf{b}^*_0 \circ \mathbf{r} - \mathbf{r})' \mathbf{R}^{-1/2} \quad 0 \right)'$, where "$\circ$" denotes the Hadamard product. We will refer to $\tilde{\mathbf{A}}^{\text{co}}$ as the *augmented weighted demeaned core*. Let $\phi_A = \phi_{\min}(\tilde{\mathbf{A}}^{\text{co}\prime}\tilde{\mathbf{A}}^{\text{co}}) \wedge 1$, [To Liangjun: I amend this definition with "$\wedge 1$" so that we avoid the ugly term $(\bar{c}+2)/\underline{c}$ in the proof of Theorem 2] which will provide an explicit solution to $\mathbf{a}^*_0$ along with an upper bound of its ℓ_1-norm.

The next lemma presents a collection of numerical properties of $\widehat{\mathbf{A}}^*$, $\widehat{\mathbf{A}}^{\text{co}}$, $\tilde{\mathbf{A}}^{\text{co}}$ and $\mathbf{a}^*_0$. Let $\underline{r} := \min_{k\in[K]} r_k$ be the smallest fraction of group size.

Lemma 2. *(a) For any $\boldsymbol{\alpha} \in \mathbb{R}^N$, we have $\|\widehat{\mathbf{A}}^* \boldsymbol{\alpha}\|_2 = \sqrt{N} \|\widehat{\mathbf{A}}^{\text{co}} \mathbf{a}\|_2 = \sqrt{N} (\mathbf{a}' \widehat{\boldsymbol{\Sigma}}^{\text{co}} (\mathbf{R} - \mathbf{r}\mathbf{r}') \widehat{\boldsymbol{\Sigma}}^{\text{co}} \mathbf{a})^{1/2}$.*

(b) $\widehat{\mathbf{A}}^{\text{co}}$ is rank deficient, $\tilde{\mathbf{A}}^{\text{co}}$ is of full column rank, and

$$\phi_A^{-1} \le 2\underline{r}^{-1} \phi^{-2}_{\min}(\widehat{\boldsymbol{\Sigma}}^{\text{co}}) + K^{-1} \phi_{\max}(\widehat{\boldsymbol{\Sigma}}^{\text{co}}) / \phi_{\min}(\widehat{\boldsymbol{\Sigma}}^{\text{co}}). \tag{22}$$

(c) $\mathbf{a}^*_0 = (\tilde{\mathbf{A}}^{\text{co}\prime}\tilde{\mathbf{A}}^{\text{co}})^{-1} \tilde{\mathbf{A}}^{\text{co}\prime} \tilde{\mathbf{b}}^{\text{co}} / N$ and $\|\mathbf{a}^*_0\|_1 \le N^{-1} \sqrt{K/\phi_A} \left(\|\mathbf{b}^*_0\|_\infty + 1 \right)$.

Part (a) of Lemma 2 characterizes the relationship between $\boldsymbol{\alpha}$ and $\mathbf{a}$ when they are premultiplied by their corresponding matrices, respectively. Part (b) concerns the ranks of $\widehat{\mathbf{A}}^{\text{co}}$ and $\tilde{\mathbf{A}}^{\text{co}}$ and the upper bound of ϕ_A^{-1}. Since $\widehat{\mathbf{A}}^{\text{co}}$ has rank $K-1$ as long as $\widehat{\boldsymbol{\Sigma}}^{\text{co}}$ has full rank K, in order to obtain an explicit expression for $\mathbf{a}_0^*$ we augment $\tilde{\mathbf{A}}^{\text{co}}$ with an additional row $\mathbf{1}_K'$ and invoke the restriction $\mathbf{1}_K'\mathbf{a}_0^* = \mathbf{1}_N'\boldsymbol{\alpha}_0^* = 0$ in the derivation. The explicit solution to $\mathbf{a}_0^*$ is given in part (c) along with its ℓ_1-norm. We are interested in the ℓ_1-norm of $\mathbf{a}_0^*$ because $\|\mathbf{a}_0^*\|_1 = \|\boldsymbol{\alpha}_0^*\|_1$ when elements in $\boldsymbol{\alpha}_0^*$ do not take opposite signs within each group, and the ℓ_1-norm of $\boldsymbol{\alpha}_0^*$ is what we penalize in the oracle dual problem (20).

While the oracle estimator has no idiosyncratic shock, the sample estimator $\widehat{\mathbf{w}}$ deviates from the oracle $\mathbf{w}_\tau^*$ due to the presence of the idiosyncratic shock and the tuning parameter τ. We will show that the effect of the idiosyncratic shock is embodied by the quantity

$$\phi_e = \|\widehat{\boldsymbol{\Sigma}}^e\|_{c2},$$

which can be viewed as an measurement of the noise level or contamination level of $\widehat{\boldsymbol{\Sigma}}^*$ in the model. Theorem 1 below reports the properties of the sample estimator.

Theorem 1. *If $\tau > \phi_e \|\mathbf{b}_0^*\|_\infty / \sqrt{N}$, then*

(a) $\|\widehat{\mathbf{w}}\|_2 \le \|\mathbf{w}_0^*\|_2 \le \|\mathbf{b}_0^*\|_\infty / \sqrt{N}$*;*

(b) $\widehat{\boldsymbol{\alpha}} \in \mathbb{S}^{\text{all}}$.

Remark 3.5. We will show that the condition $\tau > \phi_e \|\mathbf{b}_0^*\|_\infty / \sqrt{N}$ can be satisfied w.p.a.1 in the asymptotic analysis in the next section. Theorem 1(a) bounds $\|\widehat{\mathbf{w}}\|_2$, which is used in the establishment of the result in Theorem 1 (b). If the tolerance τ is sufficiently large in that it accommodates the noise, the estimator $\widehat{\boldsymbol{\alpha}}$ must have the same sign within each group. This result is proved by exploiting the KKT conditions associated with the Lagrangian of (9). The intuition is that when the specified τ is large, for any $i, j \in \mathcal{G}_k$, the difference in the noise, i.e., $\widehat{\boldsymbol{\Sigma}}^e_{\cdot i} - \widehat{\boldsymbol{\Sigma}}^e_{\cdot j}$, is unable to push the two KKT conditions to be satisfied simultaneously with $\widehat{\alpha}_i$ and $\widehat{\alpha}_j$ of different signs.

Remark 3.6. The result in Theorem 1(b) is important. It reminds us of the grouping effect of elastic net that was first proposed by Zou and Hastie (2005). A regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated regressors in the design matrix $\mathbf{X}$ tend to be equal (up to a change of sign if negatively correlated). It is well known that while Lasso yields sparse solutions in many cases, it does not have the grouping effect. In contrast, the elastic penalty, as a convex combination of the Lasso (ℓ_1) and ridge (ℓ_2) penalty, encourages the grouping effect and has the advantage of including automatically all the highly correlated variables in the group. In the presence of a latent group structure in the dominant component $\widehat{\boldsymbol{\Sigma}}^*$ of $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{\Sigma}}$ play the same role as the *regressor cross-product matrix* $\mathbf{X}'\mathbf{X}$. Then $\widehat{\boldsymbol{\Sigma}}_{i\cdot}$ and $\widehat{\boldsymbol{\Sigma}}_{j\cdot}$ are asymptotically collinear if the i-th and j-th forecasts come from the same group (e.g., $i, j \in \mathcal{G}_k$ for some $k \in [K]$), a feature similar to the highly correlated regressors in the regression framework. As a result, the ℓ_2-relaxation estimator of the weights enjoys similar properties as the elastic net estimator.

Remark 3.7. The ℓ_1 penalized form of (9) reminds us of high-dimensional Lasso estimation under sparsity. The consistency of Lasso requires that the correlation of the design matrix cannot be too high; otherwise various versions of restricted eigenvalue conditions break down (Bickel et al., 2009; Van De Geer et al., 2009; Belloni et al., 2012). When $\widehat{\mathbf{\Sigma}}$ is treated as a design matrix (or more precisely, the regressor cross-product matrix) in the regression framework, we intend to tackle highly correlated, or asymptotically perfectly collinear, design matrix. Consider the extreme case where $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}^*$ so that $\widehat{\mathbf{\Sigma}}^*$ is not contaminated by the noise component $\widehat{\mathbf{\Sigma}}^e$. If we try to solve (9) in this case, the estimated $\widehat{\boldsymbol{\alpha}}$ will be numerically very unstable due to perfect collinearity, and we cannot expect it to converge to a fixed $\boldsymbol{\alpha}_0^*$ under the ℓ_1 norm, which Lasso can often achieve in the high-dimensional regression under some restricted eigenvalue condition. Therefore we must seek a compatibility condition that is tailored for the group structure.

Lemma 3 below provides a *compatibility inequality* that links $\|\boldsymbol{\delta}\|_1$ and $\|\widehat{\mathbf{A}}\boldsymbol{\delta}\|_2^2$ for any $\boldsymbol{\delta} \in \tilde{\mathbb{S}}^{\text{all}}$. Instead of imposing a restricted eigenvalue condition as an assumption, we establish our comparability inequality under a primitive condition about ϕ_e.

Lemma 3. *If* $\phi_e \leq \frac{1}{2}\sqrt{N\phi_A/K}$, *we have* $\|\boldsymbol{\delta}\|_1 \leq 2\sqrt{K/(N\phi_A)}\|\widehat{\mathbf{A}}\boldsymbol{\delta}\|_2$ *for any* $\boldsymbol{\delta} \in \tilde{\mathbb{S}}^{\text{all}}$.

Remark 3.8. The constants $1/2$ and 2 in Lemma 3 are not important in the asymptotic analysis in Section 4. The above result means if the magnitude of the idiosyncratic shock, represented by ϕ_e, is controlled by the order $\sqrt{N\phi_A/K}$, then the ℓ_1-norm of $\boldsymbol{\delta}$ can be controlled by the ℓ_2-norm of $\|\widehat{\mathbf{A}}\boldsymbol{\delta}\|_2$ multiplied by a factor involving K/ϕ_A, which is the ratio between the number of groups K and the square of the minimal non-trivial singular value of the augmented weighted demeaned core $\tilde{\mathbf{A}}^{\text{co}}$. In the proof of Lemma 3, we introduce an original self-defined semi-norm (45) to take advantage of the group pattern explicitly.

Remark 3.9. All high-dimensional estimation problems require certain notion of sparsity to reduce dimensionality. It is helpful to compare our setting of latent group structures to the sparse Lasso. For Lasso estimation, the complexity of the problem is governed by the total number of regressors (p in Example 1) while under sparsity those non-zero coefficients control the effective number of parameters, which is assumed to be far fewer than the sample size. For the dual problem of our ℓ_2-relaxation, the complexity is the number of forecasters N whereas under the group structure the number of groups K determines the effective number of parameters, and ϕ_A is the counterpart of the restricted eigenvalue in Lasso. We will study the asymptotic behavior of ϕ_A in the asymptotic analysis in the next section.

4 Asymptotic Theory

In this section, we study the asymptotic properties of our ℓ_2-relaxed estimator of the weights. To this end, we impose some conditions and study the asymptotic properties of the estimator in the dual problem first.

4.1 Assumption

We consider a triangular array of models indexed by T and N, both of which pass to infinity. Let $\phi_{NT} = \sqrt{\frac{\log N}{T \wedge N}} \to 0$ controls the relative magnitude of N and T. Note that we allow both $N \gg T$ (as in standard high-dimensional problems) and $T \gg N$ or $T \asymp N$ in the definition of ϕ_{NT}. But we rule out the traditional case of "fixed N, large T", which can be trivially handled by (2). Furthermore, let $\boldsymbol{\Sigma}_0^e = E[\widehat{\boldsymbol{\Sigma}}^e]$, $\boldsymbol{\Delta}^e = \widehat{\boldsymbol{\Sigma}}^e - \boldsymbol{\Sigma}_0^e$, $\boldsymbol{\Sigma}_0^{\mathrm{co}} = E[\widehat{\boldsymbol{\Sigma}}^{\mathrm{co}}]$, and $\boldsymbol{\Delta}^{\mathrm{co}} = \widehat{\boldsymbol{\Sigma}}^{\mathrm{co}} - \boldsymbol{\Sigma}_0^{\mathrm{co}}$. We impose the following assumptions.

Assumption 1. *There exist universal positive finite constants $\underline{c}$ and $\overline{c}$, and C_{e0} such that*

(a) $\phi_{\max}(\boldsymbol{\Sigma}_0^e) = O(\sqrt{N}\phi_{NT})$, $\|\boldsymbol{\Delta}^e\|_\infty = O_p((T/\log N)^{-1/2})$, *and* $\|\boldsymbol{\Sigma}_0^e\|_{c2} \leq C_{e0}\phi_{\max}(\boldsymbol{\Sigma}_0^e)$*;*

(b) $\phi_{\max}(\boldsymbol{\Sigma}_0^{\mathrm{co}}) \leq \overline{c}$, $\phi_{\min}(\boldsymbol{\Sigma}_0^{\mathrm{co}}) \geq \underline{c}$, *and* $\|\boldsymbol{\Delta}^{\mathrm{co}}\|_\infty = O_p((T/\log N)^{-1/2})$.

Assumption 1(a) allows the maximal eigenvalue of the $N \times N$ matrix $\boldsymbol{\Sigma}_0^e$ to diverge to infinity, but at a limited rate $\sqrt{N}\phi_{NT}$. The sampling error of Δ_{ij}^e is controlled by $(T/\log N)^{-1/2}$ uniformly over i and j, so each element of $\widehat{\boldsymbol{\Sigma}}^e$ does not deviate too much from its population mean $\boldsymbol{\Sigma}_0^e$. The third condition in (a) is similar to but weaker than the absolute row sum condition that is frequently used to model weak cross-sectional dependence; see, e.g., Fan et al. (2013). They can be established under some low-level assumptions; see, e.g., Chapter 6 in Wainwright (2019). Assumption 1(b) is similar to (a) regarding the decomposition of the low-dimensional matrix $\widehat{\boldsymbol{\Sigma}}^{\mathrm{co}}$ into $\boldsymbol{\Sigma}_0^{\mathrm{co}}$ and $\boldsymbol{\Delta}^{\mathrm{co}}$.

Example. (Example 3, cont.) Following the notation of Example 3, we can decompose the population variance-covariance matrix $\boldsymbol{\Sigma} := E[\mathbf{e}_t\mathbf{e}_t']$ as follows: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^* + \boldsymbol{\Sigma}_0^e$, where $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}^\dagger E[\boldsymbol{\eta}_t^\dagger \boldsymbol{\eta}_t^{\dagger\prime}]\boldsymbol{\Lambda}^{\dagger\prime}$, and $\boldsymbol{\Sigma}_0^e = \boldsymbol{\Omega}_x = \{\Omega_{x,ij}\}$. The corresponding sampling error is

$$\Delta_{ij}^e = \{\mathbb{E}_T[\epsilon_{i,t}\epsilon_{j,t}] - \Omega_{x,ij}\} - \sum\nolimits_{l \in \{i,j\}} \{(\boldsymbol{\lambda}_y - \boldsymbol{\lambda}_{g_l})'\mathbb{E}_T[\boldsymbol{\eta}_t(u_{y,t+1} - u_{l,t})] + \mathbb{E}_T[u_{y,t+1}u_{l,t}]\}.$$

Then Assumption 1(a) is satisfied as long as $\|\boldsymbol{\Omega}_x\|_{\mathrm{sp}} = O(\sqrt{N}\phi_{NT})$. For the sampling error matrix, if

$$\max_{i,j \in [N]} \{|\mathbb{E}_T[\boldsymbol{\eta}_t(u_{y,t+1} - u_{i,t})]|, |\mathbb{E}_T[u_{y,t}u_{i,t}]|, |\mathbb{E}_T[u_{i,t}u_{i,j}] - \Omega_{ij,x}|\} = O_p\left((T/\log N)^{-1/2}\right),$$

then $\|\boldsymbol{\Delta}^e\|_\infty = O_p\left((T/\log N)^{-1/2}\right)$ is satisfied as well.

The next assumption imposes some restrictions on the K, τ, $\underline{r}$ and ϕ_{NT}.

4.2 Choosing the tuning parameter τ

The extent of relaxation in (8) is controlled by the tuning parameter τ. We specify the tuning parameter as

$$\tau = C_\tau K^{1/2} \phi_{NT} \tag{23}$$

for some universal constant $C_\tau > 0$. In reality K is unknown, and in the simulations and applications we will choose the constant τ based on cross-validations. When K is finite, this specification implies that τ shrinking to zero at rate ϕ_{NT}. We allow $K \to \infty$, provided the following Assumption 2 is satisfied regarding the relative magnitude of K, τ, and $\underline{r}$. Moreover, if K can be asymptotically bounded by some explicit rate function $\kappa_{N,T}$ of N and T — for example $\kappa_{N,T} = \log(N \vee T)$ — in that $K/\kappa_{N,T} \to 0$, then all the following theoretical results still hold if K is replaced by $\kappa_{N,T}$ and τ is replaced by $\tau_\kappa = C_\tau \kappa_{N,T}^{1/2} \phi_{NT}$, as long as Assumption 2 is maintained with the same replacement. [To Liangjun: I simply the specification of τ. All proofs go through.]

Assumption 2. *As $(N, T) \to \infty$,*

(a) $K^{5/2}\tau \to 0$;

(b) $\underline{r} \asymp K^{-1}$.

Assumption 2(a) sets an upper bound of the divergence of K relative to τ. Such a rate avoids the sampling error in $\widehat{\boldsymbol{\Sigma}}^e$ to offset the dominant grouping effect of the ℓ_2-relaxation in the presence of latent groups in $\widehat{\boldsymbol{\Sigma}}^*$. The particular rate $K^{5/2}\tau \asymp K^3 \phi_{NT}$ will appear as the order of convergence in Theorem 3 below. Part (b) requires that the smallest relative group size $\underline{r}$ is proportional to the reciprocal of K. Had a group included too few members, it would be difficult to assign proper weight b_{0k}^* in (17) as it is too small to matter relative to other groups.

Lemma 4 collects the implications of these assumptions.

Lemma 4. *Under Assumptions 1 and 2,*

(a) $\phi_e = O_p(\sqrt{N}\phi_{NT})$, $\phi_A^{-1} = O_p(K)$, and $\|\mathbf{b}_0^*\|_\infty = O_p\left(K^{1/2}\right)$;

(b) When τ is specified as in (23), the event $\left\{\phi_e \|\mathbf{b}_0^*\|_\infty / \sqrt{N} < \tau\right\}$ occurs w.p.a.1.

Lemma 4 (a) provides the magnitude ϕ_e, ϕ_A^{-1} and $\|\mathbf{b}_0^*\|_\infty$. Note that we allow $K \to \infty$ while $\phi_A \to 0$ and $\|\mathbf{b}_0^*\|_\infty \to \infty$ as long as the conditions in Assumption 2 are met. Part (b) shows that the key condition in the numerical properties in Theorem 1 is satisfied.

4.3 Asymptotic properties of $\widehat{\alpha}$ in the dual

We start with the dual problem (9) under Assumption 1. Following the discussion in Section 3.2, the solution to the oracle dual (20) is a set, not a singleton, due to the rank deficiency of $\boldsymbol{\Sigma}^*$. But if we want to establish asymptotic results of convergence in probability, we must declare a target to which the estimator will converge. We construct explicitly such a desirable $\boldsymbol{\alpha}_0^*$ in (24) below, denoted as $\widehat{\boldsymbol{\alpha}}^*$ where the "hat" signifies its dependence on the realization of $\widehat{\boldsymbol{\alpha}}$ and "star" indicating its validity as an oracle estimator. Define $\widehat{\boldsymbol{\alpha}}^* = (\widehat{\boldsymbol{\alpha}}^*_{\mathcal{G}_k})_{k=1}^K$, where

$$\widehat{\boldsymbol{\alpha}}^*_{\mathcal{G}_k} = a^*_{0k}\left(\frac{\widehat{\boldsymbol{\alpha}}_{\mathcal{G}_k}}{\widehat{a}_k}\cdot 1\left\{\widehat{a}_k a^*_{0k} > 0\right\} + \frac{\mathbf{1}_{N_k}}{N_k}\cdot 1\left\{\widehat{a}_k a^*_{0k} \leq 0\right\}\right), \tag{24}$$

where $\widehat{a}_k = \sum_{i\in\mathcal{G}_k}\widehat{\alpha}_i$. This $\widehat{\boldsymbol{\alpha}}^*_{\mathcal{G}_k}$is designed such that the k-th oracle group weight a^*_{0k} is distributed across group members proportionally as $\widehat{\boldsymbol{\alpha}}_{\mathcal{G}_k}/\widehat{a}_k$ when $\widehat{a}_k$ and a^*_{0k} share the same sign, whereas a^*_{0k} is distributed equally across group members when they are of the opposite signs. When $\widehat{\boldsymbol{\alpha}} \in \tilde{\mathbb{S}}^{\text{all}}$, which holds w.p.a.1 in view of Lemma 4 and Theorem 1, it is easy to verify that

$$(i)\ \widehat{\boldsymbol{\alpha}}^* \in \tilde{\mathbb{S}}^{\text{all}},\ (ii)\ \|\boldsymbol{\alpha}^*_0\|_1 = \|\widehat{\boldsymbol{\alpha}}^*\|_1 \ \text{ and } (iii)\ \widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^* \in \tilde{\mathbb{S}}^{\text{all}}. \tag{25}$$

The following theorem shows that the solution to the Lasso-type ℓ_1-penalization problem (9) is close to the desirable oracle estimator $\widehat{\boldsymbol{\alpha}}^*$.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Then*

$$\|\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*\|_1 = O_p\left(N^{-1}K^3\tau\right) \ \textit{and}\ \|\widehat{\mathbf{A}}(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*)\|_2 = O_p\left(N^{-1/2}K^2\tau\right).$$

Remark 4.1. Theorem 2 is a key result that characterizes the convergence rate of the high-dimensional parameter $\widehat{\boldsymbol{\alpha}}$ in the dual problem to its oracle group counterpart $\boldsymbol{\alpha}^*_0$, represented by the constructed unique solution $\widehat{\boldsymbol{\alpha}}^*$. Although our ultimate interest lies in the weight estimate $\widehat{\mathbf{w}}$ in the primal problem, in theoretical analysis we first work with $\widehat{\boldsymbol{\alpha}}$ in the dual problem instead. This detour is taken because the dual is an ℓ_1-penalized optimization resembles Lasso. The intensive study of Lasso in statistics and econometrics offers a set of inequalities involving the ℓ_1-norms of $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\alpha}}^*$ and their difference $(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*)$ at our disposal to analyze its asymptotic behavior. In Remark 3.9 we have given the analogy between our optimal weighting problem and the sparse regression problem. Similar to Lasso in the sparse regression, the proof counts on the compatibility condition, which we have established in Lemma 3.

4.4 Asymptotic properties of $\widehat{\mathbf{w}}$ in the primal

Given the order of $\|\mathbf{b}^*_0\|_\infty$ in Lemma 4(a), we can readily obtain

$$\|\mathbf{w}^*_0\|_1 \leq \sum_{k=1}^K |b^*_{0k}|\, r_k \leq \|\mathbf{b}^*_0\|_\infty = O_p\left(\sqrt{K}\right), \tag{26}$$

$$\|\mathbf{w}^*_0\|_2 \leq \sqrt{\sum_{k=1}^K N_k \frac{b^{*2}_{0k}}{N^2}} \leq \frac{1}{\sqrt{N}}\|\mathbf{b}^*_0\|_\infty = O_p\left(\sqrt{K/N}\right). \tag{27}$$

In particular, $\|\mathbf{w}^*_0\|_1 = O_p(1)$ and $\|\mathbf{w}^*_0\|_2 = O_p(N^{-1/2})$ in the case of fixed K. Given Theorem 2, we proceed to handle the estimator $\widehat{\mathbf{w}}$ in the ℓ_2-relaxation primal problem. Notice that for the simple average weight $\widehat{\mathbf{w}}^{sa} = \mathbf{1}_N/N$, although $\widehat{\mathbf{w}}^{sa}$ does not mimic $\mathbf{w}_0^*$ in general, the ℓ_2-distance

$$\|\widehat{\mathbf{w}}^{sa} - \mathbf{w}_0^*\|_2 \leq \|\widehat{\mathbf{w}}^{sa}\|_2 + \|\mathbf{w}_0^*\|_2 \leq (1 + \|\mathbf{b}_0^*\|_\infty)/\sqrt{N} = O_p\left(\sqrt{K/N}\right)$$

shrinks to zero in the limit. It is thus only non-trivial if we manage to show $\|\widehat{\mathbf{w}} - \mathbf{w}_0^*\|_1 = o_p(1)$ and $\|\widehat{\mathbf{w}} - \mathbf{w}_0^*\|_2 = o_p\left(N^{-1/2}\right)$, which is stated in the following Corollary 1.

Corollary 1. *Under the assumptions in Theorem 2, we have*

$$\|\widehat{\mathbf{w}} - \mathbf{w}_0^*\|_2 = O_p\left(N^{-1/2}K^2\tau\right) = o_p\left(N^{-1/2}\right) \text{ and } \|\widehat{\mathbf{w}} - \mathbf{w}_0^*\|_1 = O_p\left(K^2\tau\right) = o_p(1).$$

Remark 4.2. To connect the primal problem with the dual problem, we consider the decomposition:

$$\widehat{\mathbf{w}} - \mathbf{w}_0^* = \widehat{\mathbf{A}}(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*) + (\widehat{\mathbf{A}} - \mathbf{A}^*)\widehat{\boldsymbol{\alpha}}^*;$$

see (60) in the appendix. The ℓ_2-norm of the first term on the right-hand side is bounded by Theorem 2. Moreover, it dominates the ℓ_2-norm of the second term in which the magnitude of $\widehat{\mathbf{A}} - \mathbf{A}^*$ is well controlled under Assumption 1.

Corollary 1 establishes the meaningful convergence of the norms of the $\widehat{\mathbf{w}}$ to its oracle counterpart $\mathbf{w}_0^*$. The convergence further implies a desirable oracle inequality in Corollary 3 below, which shows that the empirical risk under $\widehat{\mathbf{w}}$ is asymptotically as small as if we knew the oracle object $\widehat{\boldsymbol{\Sigma}}^*$.

Theorem 3 (Oracle inequalities)**.** *Under the assumptions in Theorem 2, we have*

(a) $\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} \leq \mathbf{w}_0^{*\prime}\widehat{\boldsymbol{\Sigma}}^*\mathbf{w}_0^* + O_p\left(\tau K^{5/2}\right).$

Furthermore, let $\widehat{\boldsymbol{\Sigma}}^{\text{new}}$ *and* $\widehat{\boldsymbol{\Sigma}}^{*\text{new}}$ *be the counterparts of* $\widehat{\boldsymbol{\Sigma}}$ *and* $\widehat{\boldsymbol{\Sigma}}^*$ *from a new (testing) sample, which can be dependent or independent of the training dataset used to estimate* $\widehat{\mathbf{w}}$ *and* $\mathbf{w}_0^*$*. If the testing dataset is obtained from the same data generating process as that of the training dataset,*

(b) $\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}^{\text{new}}\widehat{\mathbf{w}} \leq \mathbf{w}_0^{*\prime}\widehat{\boldsymbol{\Sigma}}^{*\text{new}}\mathbf{w}_0^* + O_p\left(\tau K^{5/2}\right);$

(c) Let $Q(\boldsymbol{\Sigma}_0)$ *be the minimum of* $\min_{\mathbf{w}\in\mathbb{R}^N, \mathbf{w}'\mathbf{1}_N = 1} \mathbf{w}'\boldsymbol{\Sigma}_0\mathbf{w}$*, where* $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_0^* + \boldsymbol{\Sigma}_0^e$*. Then*

$$\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} \leq \widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}^{\text{new}}\widehat{\mathbf{w}} \leq Q(\boldsymbol{\Sigma}_0) + O_p\left(\tau K^{5/2}\right).$$

Remark 4.3. Theorem 3(a) is an in-sample oracle inequality, and (b) is an out-of-sample oracle inequality. The proof is a term-by-term analysis of the difference between $\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}$ and $\mathbf{w}_0^{*\prime}\widehat{\boldsymbol{\Sigma}}^*\mathbf{w}_0^*$ caused by the idiosyncratic shock. Again, because the magnitude of the idiosyncratic shock is controlled by Assumption 1, the convergence of the weight in Corollary 1 allows the sample risk $\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}$ to approximate the oracle risk $\mathbf{w}_0^{*\prime}\widehat{\boldsymbol{\Sigma}}^*\mathbf{w}_0^*$. The approximation is nontrivial by noting that $\mathbf{w}_0^{*\prime}\widehat{\boldsymbol{\Sigma}}^*\mathbf{w}_0^*$ and $\mathbf{w}_0^{*\prime}\widehat{\boldsymbol{\Sigma}}^{*\text{new}}\mathbf{w}_0^*$ are typically bounded away from 0 given the low rank structure of $\widehat{\boldsymbol{\Sigma}}^{*\text{new}}$ and that $\tau K^{5/2} \to 0$ under Assumption 2(a). In the other words, under appropriate conditions, the risk of our sample estimator is as low as if we knew the oracle group membership which is infeasible in reality, up to an asymptotically negligible term.

Remark 4.4: While our ℓ_2-relaxation regularizes the combination weights, there is another line of literature of regularizing the high-dimensional VC estimation or the its inverse (the precision matrix); see Bickel and Levina (2008), Fan et al. (2013), and the overview Fan et al. (2016). Theorem 3(c) implies that the in-sample and out-of-sample risks coming out the ℓ_2-relaxation is competitive with the resultant risk from this approach. Specifically, $Q(\mathbf{\Sigma}_0)$ is Bates and Granger (1969)'s optimal risk under the population VC $\mathbf{\Sigma}_0$, which is the target of high-dimensional VC estimation or precision matrix estimation. Compared to the oracle with respect to the idiosyncratic-noise-free $\widehat{\mathbf{\Sigma}}^*$, the population VC $\mathbf{\Sigma}_0$ takes account both the low-dimensional $\mathbf{\Sigma}_0^*$ and $\mathbf{\Sigma}_0^e$. Even if $\mathbf{\Sigma}_0$ can be estimated so well such that the estimation error is negligible, our out-of-sample risk $\widehat{\mathbf{w}}'\widehat{\mathbf{\Sigma}}^{\text{new}}\widehat{\mathbf{w}}$ is within an $O_p\left(\tau K^{5/2}\right)$ tolerance level.

In summary, we have shown that under the high-dimensional asymptotic framework where $N/T \to \infty$ is allowed as $(N,T) \to \infty$, we can construct a unique oracle $\widehat{\boldsymbol{\alpha}}^*$ that satisfies a set of desirable properties. While the dual problem is an ℓ_1-penalized optimization problem, we establish in Theorem 1 the convergence of $\widehat{\boldsymbol{\alpha}}$ to this oracle estimator by borrowing some techniques that deal with the ℓ_1-regularization, thanks to the amenable comparability condition under Assumption 1(d). Then Theorem 1 can be extended to the convergence of the weight $\widehat{\mathbf{w}}$ in Corollary 1 and furthermore the convergence of the sample risk to the oracle risk in Theorem 3.

5 Conclusion

We propose a new machine learning algorithm for forecast combination. We justify the theoretical appeal under the group structure...

we develop the asymptotic theory. We conduct extensive simulations. We demonstrate the performance of our estimator in two real-data examples, one from microeconomics and the other from macroeconomics.

Additional forms of penalty can used to accompany ℓ_2 relaxation. For example, if sparsity is desirable, we may consider adding another constraint $\|\mathbf{w}\|_1 \leq \tau_1$ for some tuning parameter τ_1, which is similar to the idea of mixed ℓ_2-ℓ_1 penalty of elastic net (Zou and Hastie, 2005) If non-negative weights are desirable, we can add constraints $w_i \geq 0$ for all i, similar to Jagannathan and Ma (2003). The idea of ℓ_2-relaxation can also be applied to portfolio optimization problems.

be In cases when extra properties of the combination weights

References

Ao, M., L. Yingying, and X. Zheng (2019). Approaching mean-variance efficiency for large portfolios. The Review of Financial Studies 32(7), 2890–2919.

Bates, J. M. and C. W. Granger (1969). The combination of forecasts. Operational Research Quarterly, 451–468.

Bayer, S. (2018). Combining value-at-risk forecasts using penalized quantile regressions. Econometrics and Statistics 8, 56–77.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics 37(4), 1705–1732.

Bickel, P. J. and E. Levina (2008). Regularized estimation of large covariance matrices. The Annals of Statistics 36(1), 199–227.

Bonhomme, S., T. Lamadon, and E. Manresa (2017). Discretizing unobserved heterogeneity. University of Chicago, Becker Friedman Institute for Economics Working Paper (2019-16).

Bonhomme, S. and E. Manresa (2015). Grouped patterns of heterogeneity in panel data. Econometrica 83(3), 1147–1184.

Boyd, S. and L. Vandenberghe (2004). Convex optimization. Cambridge University Press.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics 35(6), 2313–2351.

Chan, F. and L. L. Pauwels (2018). Some theoretical results on forecast combinations. International Journal of Forecasting 34(1), 64–74.

Claeskens, G., J. R. Magnus, A. L. Vasnev, and W. Wang (2016). The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting 32(3), 754–762.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. International Journal of forecasting 5(4), 559–583.

Conflitti, C., C. De Mol, and D. Giannone (2015). Optimal combination of survey forecasts. International Journal of Forecasting 31(4), 1096–1103.

Coulombe, P. G., M. Leroux, D. Stevanovic, S. Surprenant, et al. (2020). How is machine learning useful for macroeconomic forecasting? Technical report, CIRANO.

Diebold, F. X. and M. Shin (2019). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. International Journal of Forecasting 35(4), 1679–1691.

Disatnik, D. and S. Katz (2012). Portfolio optimization using a block structure for the covariance matrix. Journal of Business Finance & Accounting 39(5-6), 806–843.

Elliott, G., A. Gargano, and A. Timmermann (2013). Complete subset regressions. Journal of Econometrics 177(2), 357–373.

Engle, R. and B. Kelly (2012). Dynamic equicorrelation. Journal of Business & Economic Statistics 30(2), 212–228.

Fan, J., A. Furger, and D. Xiu (2016). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. Journal of Business & Economic Statistics 34(4), 489–503.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96(456), 1348–1360.

Fan, J., Y. Liao, and H. Liu (2016). An overview of the estimation of large covariance and precision matrices. The Econometrics Journal 19(1), C1–C32.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75(4), 603–680.

Fan, J., J. Zhang, and K. Yu (2012). Vast portfolio selection with gross-exposure constraints. Journal of the American Statistical Association 107(498), 592–606.

Granger, C. W. and R. Ramanathan (1984). Improved methods of combining forecasts. Journal of Forecasting 3(2), 197–204.

Hsiao, C. and S. K. Wan (2014). Is there an optimal forecast combination? Journal of Econometrics 178, 294–309.

Jagannathan, R. and T. Ma (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. The Journal of Finance 58(4), 1651–1683.

Konzen, E. and F. A. Ziegelmann (2016). Lasso-type penalties for covariate selection and forecasting in time series. Journal of Forecasting 35(7), 592–612.

Kotchoni, R., M. Leroux, and D. Stevanovic (2019). Macroeconomic forecast accuracy in a data-rich environment. Journal of Applied Econometrics 34(7), 1050–1072.

Ledoit, O. and M. Wolf (2004). Honey, i shrunk the sample covariance matrix. The Journal of Portfolio Management 30(4), 110–119.

Ledoit, O. and M. Wolf (2017). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. The Review of Financial Studies 30(12), 4349–4388.

Li, J. and W. Chen (2014). Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. International Journal of Forecasting 30(4), 996–1015.

Roccazzella, F., P. Gambetti, and F. Vrins (2020). Optimal and robust combination of forecasts via constrained optimization and shrinkage. Technical report, LFIN Working Paper Series, 2020/6, 1–2.

Shi, Z. (2016). Econometric estimation with high-dimensional moment equalities. Journal of Econometrics 195(1), 104–119.

Smith, J. and K. F. Wallis (2009). A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics 71(3), 331–355.

Stasinakis, C., G. Sermpinis, K. Theofilatos, and A. Karathanasopoulos (2016). Forecasting us unemployment with radial basis neural networks, kalman filters and support vector regressions. Computational Economics 47(4), 569–587.

Stock, J. H. and M. W. Watson (2004). Combination forecasts of output growth in a seven-country data set. Journal of Forecasting 23(6), 405–430.

Su, L. and G. Ju (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. Journal of Econometrics 206(2), 554–573.

Su, L., Z. Shi, and P. C. Phillips (2016). Identifying latent structures in panel data. Econometrica 84(6), 2215–2264.

Su, L., X. Wang, and S. Jin (2019). Sieve estimation of time-varying panel data models with latent structures. Journal of Business & Economic Statistics 37(2), 334–349.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288.

Tikhonov (1977). Solutions of Ill-Posed Problems. Winston and Sons, Washington, DC.

Timmermann, A. (2006). Forecast combinations. Handbook of Economic Forecasting 1, 135–196.

Van De Geer, S. A., P. Bühlmann, et al. (2009). On the conditions used to prove oracle results for the lasso. Electronic Journal of Statistics 3, 1360–1392.

Vogt, M. and O. Linton (2017). Classification of non-parametric regression functions in longitudinal data models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79(1), 5–27.

Vogt, M. and O. Linton (2020). Multiscale clustering of nonparametric regression curves. Journal of Econometrics.

Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint, Volume 48. Cambridge University Press.

Wang, W. and L. Su (2020). Identifying latent group structures in nonlinear panels. Journal of Econometrics.

Wilms, I., J. Rombouts, and C. Croux (2018). Multivariate lasso-based forecast combinations for stock market volatility.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: series B (Statistical Methodology) 67(2), 301–320.

Appendix

This appendix is composed of two sections. Appendix A contains the proofs of the main results in the paper. Appendix B contains some additional simulation and empirical examples.

A Proofs of the Main Results

A.1 Proof of the Results in Section 2

Proof of Lemma 1. First, we can rewrite the minimization problem in (8) in terms of linear constraints:

$$\min_{(\mathbf{w},\gamma)\in\mathbb{R}^{N+1}} \frac{1}{2}\left\|\mathbf{w}\right\|_2^2$$

$$\text{s.t. } \mathbf{w}'\mathbf{1}_N - 1 = 0,\ \begin{pmatrix}\widehat{\boldsymbol{\Sigma}} & \mathbf{1}_N\end{pmatrix}\begin{pmatrix}\mathbf{w}' & \gamma'\end{pmatrix}' \leq \tau\mathbf{1}_N,\ \text{and } -\begin{pmatrix}\widehat{\boldsymbol{\Sigma}} & \mathbf{1}_N\end{pmatrix}\begin{pmatrix}\mathbf{w}' & \gamma'\end{pmatrix}' \leq \tau\mathbf{1}_N \qquad (28)$$

where "≤" holds elementwise hereafter. Define the Lagrangian function as

$$\begin{aligned}\mathcal{L}\left(\mathbf{w},\gamma;\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\alpha_3\right) &= \frac{1}{2}\mathbf{w}'\mathbf{w} + \boldsymbol{\alpha}_1'\left(\begin{pmatrix}\widehat{\boldsymbol{\Sigma}} & \mathbf{1}_N\end{pmatrix}\begin{pmatrix}\mathbf{w}\\ \gamma\end{pmatrix} - \tau\mathbf{1}_N\right)\\ &\quad -\boldsymbol{\alpha}_2'\left(\begin{pmatrix}\widehat{\boldsymbol{\Sigma}} & \mathbf{1}_N\end{pmatrix}\begin{pmatrix}\mathbf{w}\\ \gamma\end{pmatrix} + \tau\mathbf{1}_N\right) + \alpha_3\left(\mathbf{w}'\mathbf{1}_N - 1\right) \qquad (29)\end{aligned}$$

and the associated Lagrangian dual function as $g\left(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\alpha_3\right) = \inf_{\mathbf{w},\gamma}\mathcal{L}\left(\mathbf{w},\gamma;\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\alpha_3\right),$ where $\boldsymbol{\alpha}_1 \geq 0$, $\boldsymbol{\alpha}_2 \geq 0$, and α_3 are the Lagrangian multipliers for the three constraints, respectively.

Let $\varphi\left(\mathbf{w},\gamma\right) = \frac{1}{2}\left\|\mathbf{w}\right\|_2^2$, the objective function in (28). Define its conjugate function as

$$\varphi^*\left(\mathbf{a},b\right) = \sup_{\mathbf{w},\gamma}\left\{\mathbf{a}'\mathbf{w} + b\gamma - \frac{1}{2}\left\|\mathbf{w}\right\|_2^2\right\} = \begin{cases}\frac{1}{2}\left\|\mathbf{a}\right\|_2^2 & \text{if } b = 0\\ \infty & \text{otherwise}\end{cases}.$$

The linear constraints indicate an explicit dual function (See Boyd and Vandenberghe, 2004, p.221):

$$\begin{aligned}& g\left(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\alpha_3\right)\\ =\ & -\tau\mathbf{1}_N'\left(\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2\right) - \alpha_3 - \varphi^*\left(\widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{\alpha}_2-\boldsymbol{\alpha}_1\right) - \alpha_3\mathbf{1}_N,\ \mathbf{1}_N'\left(\boldsymbol{\alpha}_2-\boldsymbol{\alpha}_1\right)\right)\\ =\ & \begin{cases}-\tau\mathbf{1}_N'\left(\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2\right) - \alpha_3 - \frac{1}{2}\left\|\widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{\alpha}_2-\boldsymbol{\alpha}_1\right) - \alpha_3\mathbf{1}_N\right\|_2^2 & \text{if } \mathbf{1}_N'\left(\boldsymbol{\alpha}_2-\boldsymbol{\alpha}_1\right) = 0\\ \infty & \text{otherwise}\end{cases}.\end{aligned}$$

Let $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1$. When $\tau > 0$, the two inequalities $\widehat{\boldsymbol{\Sigma}}_{i\cdot}\mathbf{w} + \gamma \leq \tau$ and $-\widehat{\boldsymbol{\Sigma}}_{i\cdot}\mathbf{w} - \gamma \leq \tau$ cannot be binding simultaneously. The associated Lagrangian multipliers α_{1i} and α_{2i} must satisfy $\alpha_{1i}\cdot\alpha_{2i} = 0$ for all $i \in [N]$. This implies that $\left\|\boldsymbol{\alpha}\right\|_1 = \mathbf{1}_N'\boldsymbol{\alpha}_1 + \mathbf{1}_N'\boldsymbol{\alpha}_2$ so that the dual problem can be simplified as

$$\max_{\boldsymbol{\alpha},\alpha_3} \left\{ -\frac{1}{2} \left\| \widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} - \alpha_3 \mathbf{1}_N \right\|_2^2 - \alpha_3 - \tau \left\| \boldsymbol{\alpha} \right\|_1 \right\} \quad \text{s.t. } \mathbf{1}_N' \boldsymbol{\alpha} = 0. \tag{30}$$

Taking the partial derivative of the above criterion function with respect to α_3 yields

$$(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} - \alpha_3 \mathbf{1}_N)' \mathbf{1}_N - 1 = 0,$$

or equivalently, $\alpha_3 = \frac{1}{N}(\mathbf{1}_N' \widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} - 1)$. Then

$$\left\| \widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} - \alpha_3 \mathbf{1}_N \right\|_2^2 = \left\| \widehat{\mathbf{A}}\boldsymbol{\alpha} + \frac{\mathbf{1}_N}{N} \right\|_2^2 = \boldsymbol{\alpha}' \widehat{\mathbf{A}}' \widehat{\mathbf{A}} \boldsymbol{\alpha} + \frac{1}{N}$$

where $\widehat{\mathbf{A}} = \left(\mathbf{I}_N - N^{-1} \mathbf{1}_N \mathbf{1}_N' \right) \widehat{\boldsymbol{\Sigma}}$ as defined in the main text. We conclude that the dual problem (30) is equivalent to

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \left\{ \frac{1}{2} \boldsymbol{\alpha}' \widehat{\mathbf{A}}' \widehat{\mathbf{A}} \boldsymbol{\alpha} + \frac{1}{N} \mathbf{1}_N' \widehat{\boldsymbol{\Sigma}} \boldsymbol{\alpha} + \tau \left\| \boldsymbol{\alpha} \right\|_1 - \frac{1}{2N} \right\} \text{ subject to } \mathbf{1}_N' \boldsymbol{\alpha} = 0, \tag{31}$$

where we keep the constant $-\frac{1}{2N}$ which is irrelevant to the optimization.

When $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_2 - \widehat{\boldsymbol{\alpha}}_1$ is the solution to (31), the solution of α_3 in (30) is $\widehat{\alpha}_3 = \frac{1}{N}(\mathbf{1}_N' \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\alpha}} - 1)$. The first order condition of (29) with respect to $\mathbf{w}$ evaluated at the solution gives

$$\mathbf{0}_N = \widehat{\mathbf{w}} + \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\alpha}}_2) + \widehat{\alpha}_3 \mathbf{1}_N = \widehat{\mathbf{w}} - \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}} + \frac{1}{N}(\mathbf{1}_N' \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\alpha}} - 1)\mathbf{1}_N = \widehat{\mathbf{w}} - \widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}} - \frac{1}{N}\mathbf{1}_N.$$

as $\mathbf{1}_N' \widehat{\boldsymbol{\alpha}} = 0$. The result in (10) follows.■

A.2 Proofs of the Results in Section 3

Proof of Lemma 2. Part (a): Note that

$$\left\| \widehat{\mathbf{A}}^* \boldsymbol{\alpha} \right\|_2^2 = \boldsymbol{\alpha}' \widehat{\boldsymbol{\Sigma}}^* \left(\mathbf{I}_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N' \right) \widehat{\boldsymbol{\Sigma}}^* \boldsymbol{\alpha} = \boldsymbol{\alpha}' \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{\Sigma}}^* \boldsymbol{\alpha} - \frac{1}{N} \left(\mathbf{1}_N' \widehat{\boldsymbol{\Sigma}}^* \boldsymbol{\alpha} \right)^2. \tag{32}$$

The group structure in $\widehat{\boldsymbol{\Sigma}}^*$ implies $\widehat{\boldsymbol{\Sigma}}^* \boldsymbol{\alpha} = \left(\widehat{\boldsymbol{\Sigma}}_{1\cdot}^{\text{co}} \mathbf{a} \cdot \mathbf{1}_{N_1}', \ldots, \widehat{\boldsymbol{\Sigma}}_{K\cdot}^{\text{co}} \mathbf{a} \cdot \mathbf{1}_{N_K}' \right)'$. Therefore, we have

$$\boldsymbol{\alpha}' \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{\Sigma}}^* \boldsymbol{\alpha} = \sum_{k=1}^K N_k \left(\widehat{\boldsymbol{\Sigma}}_{k\cdot}^{\text{co}} \mathbf{a} \right)^2 = N \sum_{k=1}^K r_k \left(\widehat{\boldsymbol{\Sigma}}_{k\cdot}^{\text{co}} \mathbf{a} \right)^2 = N \mathbf{a}' \widehat{\boldsymbol{\Sigma}}^{\text{co}} \mathbf{R} \widehat{\boldsymbol{\Sigma}}^{\text{co}} \mathbf{a},$$
$$\mathbf{1}_N' \widehat{\boldsymbol{\Sigma}}^* \boldsymbol{\alpha} = \sum_{k=1}^K N_k \widehat{\boldsymbol{\Sigma}}_{k\cdot}^{\text{co}} \mathbf{a} = N \sum_{k=1}^K r_k \widehat{\boldsymbol{\Sigma}}_{k\cdot}^{\text{co}} \mathbf{a} = N \mathbf{r}' \widehat{\boldsymbol{\Sigma}}^{\text{co}} \mathbf{a}.$$

Substituting these two equations to (32) yields

$$\left\| \widehat{\mathbf{A}}^* \boldsymbol{\alpha} \right\|_2^2 = N \mathbf{a}' \widehat{\boldsymbol{\Sigma}}^{\text{co}} \mathbf{R} \widehat{\boldsymbol{\Sigma}}^{\text{co}} \mathbf{a} - N \left(\mathbf{r}' \widehat{\boldsymbol{\Sigma}}^{\text{co}} \mathbf{a} \right)^2 = N \mathbf{a}' \widehat{\boldsymbol{\Sigma}}^{\text{co}} \left(\mathbf{R} - \mathbf{r}\mathbf{f}' \right) \widehat{\boldsymbol{\Sigma}}^{\text{co}} \mathbf{a}.$$

On the other hand, noticing $\mathbf{R}\mathbf{1}_K = \mathbf{r}$ and $\mathbf{1}_K'\mathbf{R}\mathbf{1}_K = \mathbf{1}_K'\mathbf{r} = 1$, we have

$$\begin{aligned}
N\left\|\widehat{\mathbf{A}}^{\text{co}}\mathbf{a}\right\|_2^2 &= N\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{a}'\left(\mathbf{I}_K - \mathbf{r}\cdot\mathbf{1}_K'\right)\mathbf{R}\left(\mathbf{I}_K - \mathbf{1}_K\cdot\mathbf{r}'\right)\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{a} \\
&= N\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{a}'\left(\mathbf{R} - \mathbf{R}\mathbf{1}_K\cdot\mathbf{r}' - \mathbf{r}\cdot\mathbf{1}_K'\mathbf{R} + \mathbf{r}\cdot\mathbf{1}_K'\mathbf{R}\mathbf{1}_K\cdot\mathbf{r}'\right)\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{a} \\
&= N\mathbf{a}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}\left(\mathbf{R} - \mathbf{r}\mathbf{r}'\right)\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{a}.
\end{aligned}$$

The two sides are equal.

Part (b): Notice that $\mathbf{I}_K - \mathbf{r}\cdot\mathbf{1}_K'$ is idempotent. Since $\mathbf{R}^{1/2}$ and $\widehat{\boldsymbol{\Sigma}}^{\text{co}}$ are both of full rank, the $K\times K$ matrix $\widehat{\mathbf{A}}^{\text{co}}$ has its rank equal to $\text{rank}(\mathbf{I}_K - \mathbf{r}\cdot\mathbf{1}_K') = \text{trace}(\mathbf{I}_K - \mathbf{r}\cdot\mathbf{1}_K') = K-1$. In other words, $\widehat{\mathbf{A}}^{\text{co}}$ is rank deficient and its null space is one-dimensional.

It is easy to verify that the null space of $\widehat{\mathbf{A}}^{\text{co}}$ is $\ker(\widehat{\mathbf{A}}^{\text{co}}) = \left\{c(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K : c\in\mathbb{R}\backslash\{0\}\right\}$, as $\widehat{\mathbf{A}}^{\text{co}}\times c(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K = c\mathbf{R}^{1/2}\left(\mathbf{I}_K - \mathbf{1}_K\cdot\mathbf{r}'\right)\mathbf{1}_K = \mathbf{0}_N$, and $\mathbf{1}_K \notin \ker(\widehat{\mathbf{A}}^{\text{co}})$. Moreover, since $\mathbf{1}_K'\times c(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K = c\mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K \neq 0$, by construction $\tilde{\mathbf{A}}^{\text{co}}$ is of full column rank as the additional row $\mathbf{1}_K'$ is not in the null space of $\widehat{\mathbf{A}}^{\text{co}}$.

Next, note that

$$\begin{aligned}
\tilde{\mathbf{A}}^{\text{co}\prime}\tilde{\mathbf{A}}^{\text{co}} &= \widehat{\mathbf{A}}^{\text{co}\prime}\widehat{\mathbf{A}}^{\text{co}} + \mathbf{1}_K\mathbf{1}_K' = \widehat{\boldsymbol{\Sigma}}^{\text{co}}\left(\mathbf{I}_K - \mathbf{r}\cdot\mathbf{1}_K'\right)\mathbf{R}\left(\mathbf{I}_K - \mathbf{1}_K\cdot\mathbf{r}'\right)\widehat{\boldsymbol{\Sigma}}^{\text{co}} + \mathbf{1}_K\mathbf{1}_K' \\
&= \widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{R}\widehat{\boldsymbol{\Sigma}}^{\text{co}} + \mathbf{1}_K\mathbf{1}_K' - \widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{r}\mathbf{r}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}.
\end{aligned}$$

By the Sherman-Morrison formula (80),

$$(\tilde{\mathbf{A}}^{\text{co}\prime}\tilde{\mathbf{A}}^{\text{co}})^{-1} = \mathbf{A}_1^{-1} + \frac{\mathbf{A}_1^{-1}\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{r}\mathbf{r}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{A}_1^{-1}}{1 - \mathbf{r}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{A}_1^{-1}\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{r}}, \tag{33}$$

where $\mathbf{A}_1 = \widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{R}\widehat{\boldsymbol{\Sigma}}^{\text{co}} + \mathbf{1}_K\mathbf{1}_K'$ and moreover

$$\mathbf{A}_1^{-1} = \mathbf{A}_2^{-1} - \frac{\mathbf{A}_2^{-1}\mathbf{1}_K\mathbf{1}_K'\mathbf{A}_2^{-1}}{1 + \mathbf{1}_K'\mathbf{A}_2^{-1}\mathbf{1}_K} \tag{34}$$

by (79), where $\mathbf{A}_2 = \widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{R}\widehat{\boldsymbol{\Sigma}}^{\text{co}}$. Obviously $\phi_{\min}\left(\mathbf{A}_2\right) \geq \underline{r}\phi_{\min}^2(\widehat{\boldsymbol{\Sigma}}^{\text{co}})$.

The denominator of the second term on the right-hand side of (33) is

$$\begin{aligned}
1 - \mathbf{r}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{A}_1^{-1}\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{r} &= 1 - \mathbf{r}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}\left(\mathbf{A}_2^{-1} - \frac{\mathbf{A}_2^{-1}\mathbf{1}_K\mathbf{1}_K'\mathbf{A}_2^{-1}}{1 + \mathbf{1}_K'\mathbf{A}_2^{-1}\mathbf{1}_K}\right)\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{r} \\
&= \mathbf{r}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}\frac{\mathbf{A}_2^{-1}\mathbf{1}_K\mathbf{1}_K'\mathbf{A}_2^{-1}}{1 + \mathbf{1}_K'\mathbf{A}_2^{-1}\mathbf{1}_K}\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{r} = \frac{\left[\mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K\right]^2}{1 + \mathbf{1}_K'\mathbf{A}_2^{-1}\mathbf{1}_K} > 0,
\end{aligned} \tag{35}$$

where the second equality follows by $\mathbf{r}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{A}_2^{-1}\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{r} = \mathbf{r}'\mathbf{R}^{-1}\mathbf{r} = \mathbf{1}_K'\mathbf{r} = 1$, and the third equality by $\mathbf{r}'\widehat{\boldsymbol{\Sigma}}^{\text{co}}\mathbf{A}_2^{-1}\mathbf{1}_K = \mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K$. The numerator of the second term on the right-hand side of (33) has rank 1, and thus

$$\begin{aligned}\phi_{\max}\left(\mathbf{A}_1^{-1}\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{r}\mathbf{r}'\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{A}_1^{-1}\right) &= \text{trace}\left(\mathbf{A}_1^{-1}\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{r}\mathbf{r}'\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{A}_1^{-1}\right) = \mathbf{r}'\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{A}_1^{-2}\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{r} \\ &\leq \mathbf{r}'\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{A}_2^{-2}\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{r} = \mathbf{r}'\mathbf{R}^{-1}(\widehat{\mathbf{\Sigma}}^{\text{co}})^{-2}\mathbf{R}^{-1}\mathbf{r} = \mathbf{1}_K'(\widehat{\mathbf{\Sigma}}^{\text{co}})^{-2}\mathbf{1}_K. \quad (36)\end{aligned}$$

Apply the spectral norm to (33):

$$\begin{aligned}\phi_A^{-1} &= \|(\tilde{\mathbf{A}}^{\text{co}\prime}\tilde{\mathbf{A}}^{\text{co}})^{-1}\|_{\text{sp}} \\ &\leq \|\mathbf{A}_1^{-1}\|_{\text{sp}} + \frac{\|\mathbf{A}_1^{-1}\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{r}\mathbf{r}'\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{A}_1^{-1}\|_{\text{sp}}}{1-\mathbf{r}'\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{A}_1^{-1}\widehat{\mathbf{\Sigma}}^{\text{co}}\mathbf{r}} \\ &\leq \|\mathbf{A}_2^{-1}\|_{\text{sp}} + \mathbf{1}_K'(\widehat{\mathbf{\Sigma}}^{\text{co}})^{-2}\mathbf{1}_K \times \frac{1+\mathbf{1}_K'\mathbf{A}_2^{-1}\mathbf{1}_K}{\left[\mathbf{1}_K'(\widehat{\mathbf{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K\right]^2} \\ &\leq \underline{r}^{-1}\phi_{\text{min}}^{-2}(\widehat{\mathbf{\Sigma}}^{\text{co}}) + \frac{\phi_{\text{min}}^{-1}(\widehat{\mathbf{\Sigma}}^{\text{co}})}{\mathbf{1}_K'(\widehat{\mathbf{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K}\left(1+\underline{r}^{-1}\phi_{\text{min}}^{-1}(\widehat{\mathbf{\Sigma}}^{\text{co}})\cdot\mathbf{1}_K'(\widehat{\mathbf{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K\right) \\ &\leq \underline{r}^{-1}\phi_{\text{min}}^{-2}(\widehat{\mathbf{\Sigma}}^{\text{co}}) + \frac{\phi_{\text{max}}(\widehat{\mathbf{\Sigma}}^{\text{co}})}{K\phi_{\text{min}}(\widehat{\mathbf{\Sigma}}^{\text{co}})} + \underline{r}^{-1}\phi_{\text{min}}^{-2}(\widehat{\mathbf{\Sigma}}^{\text{co}})\end{aligned}$$

where the first inequality follows by collecting (34), (35) and (36).

Part (c): Setting $\tau = 0$ in (21), we have

$$\widehat{\mathbf{A}}^*\boldsymbol{\alpha}_0^* = \mathbf{w}_0^* - \frac{\mathbf{1}_N}{N}. \quad (37)$$

Premultiplying both sides of the above equation by the $K\times N$ block diagonal matrix diag$(r_1^{-1/2}\mathbf{1}_{N_1}', ...,$ $r_K^{-1/2}\mathbf{1}_{N_K}')$, we obtain

$$N\widehat{\mathbf{A}}^{\text{co}}\mathbf{a}_0^* = \mathbf{R}^{-1/2}\left(\mathbf{b}_0^*\circ\mathbf{r}-\mathbf{r}\right). \quad (38)$$

As $\widehat{\mathbf{A}}^{\text{co}}$ is part of the extended matrix $\tilde{\mathbf{A}}^{\text{co}}$, the above equation implies

$$N\tilde{\mathbf{A}}^{\text{co}}\mathbf{a}_0^* = \begin{pmatrix} N\widehat{\mathbf{A}}^{\text{co}}\mathbf{a}_0^* \\ N\mathbf{1}_K'\mathbf{a}_0^* \end{pmatrix} = \begin{pmatrix} \mathbf{R}^{-1/2}\left(\mathbf{b}_0^*\circ\mathbf{r}-\mathbf{r}\right) \\ 0 \end{pmatrix} = \tilde{\mathbf{b}}^{\text{co}},$$

where we use the fact that $N\mathbf{1}_K^{'}\mathbf{a}_0^* = N\mathbf{1}_N^{'}\boldsymbol{\alpha}_0^* = 0$. Since $\tilde{\mathbf{A}}^{\text{co}}$ is of full column rank, we explicitly solve $\mathbf{a}_0^* = \left(\tilde{\mathbf{A}}^{\text{co}\prime}\tilde{\mathbf{A}}^{\text{co}}\right)^{-1}\tilde{\mathbf{A}}^{\text{co}\prime}\tilde{\mathbf{b}}^{\text{co}}/N$. Its ℓ_2-norm is bounded by

$$\begin{aligned}\|\mathbf{a}_0^*\|_2 &\leq \|\left(\tilde{\mathbf{A}}^{\text{co}\prime}\tilde{\mathbf{A}}^{\text{co}}\right)^{-1}\tilde{\mathbf{A}}^{\text{co}\prime}\|_{\text{sp}}\|\tilde{\mathbf{b}}^{\text{co}}\|_2/N \\ &\leq \frac{1}{N\sqrt{\phi_A}}\left(\|\mathbf{b}_0^*\circ\mathbf{r}^{1/2}\|_2+\|\mathbf{r}^{1/2}\|_2\right) \leq \frac{1}{N\sqrt{\phi_A}}\left(\|\mathbf{b}_0^*\|_\infty+1\right), \quad (39)\end{aligned}$$

where $\mathbf{r}^{1/2} = (r_1^{1/2}, \ldots, r_k^{1/2})'$. In addition, the Cauchy-Schwarz inequality entails

$$\|\mathbf{a}_0^*\|_1 \leq \sqrt{K}\,\|\mathbf{a}_0^*\|_2 = \frac{1}{N}\sqrt{\frac{K}{\phi_A}}\,(\|\mathbf{b}_0^*\|_\infty + 1) \tag{40}$$

as stated in the lemma. ■

Proof of Theorem 1. **Part (a)**: Substituting $\mathbf{w}_0^*$ and γ_0^* into the constraint in (8), we obtain

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Sigma}}\mathbf{w}_0^* + \gamma_0^*\mathbf{1}_N\right\|_\infty &= \|\widehat{\boldsymbol{\Sigma}}^*\mathbf{w}_0^* + \gamma_0^*\mathbf{1}_N + \widehat{\boldsymbol{\Sigma}}^e\mathbf{w}_0^*\|_\infty = \|\widehat{\boldsymbol{\Sigma}}^e\mathbf{w}_0^*\|_\infty \\ &= \max_i \|\widehat{\boldsymbol{\Sigma}}^e_{i\cdot}\mathbf{w}_0^*\|_\infty \leq \|\widehat{\boldsymbol{\Sigma}}^e\|_{c2}\|\mathbf{w}_0^*\|_2 \leq \phi_e\|\mathbf{b}_0^*\|_\infty/\sqrt{N}.\end{aligned} \tag{41}$$

where the second equality follows by the KKT condition $\widehat{\boldsymbol{\Sigma}}^*\mathbf{w}_0^* + \gamma_0^*\mathbf{1}_N = 0$. The presumption $\phi_e\,\|\mathbf{b}_0^*\|_\infty/\sqrt{N} < \tau$ in the statement makes sure that $\|\widehat{\boldsymbol{\Sigma}}\mathbf{w}_0^* + \gamma_0^*\mathbf{1}_N\|_\infty < \tau$ holds with strict inequality. This inequality means that $(\mathbf{w}_0^*, \gamma_0^*)$ lies in the interior of the feasible set of (8). Because $\widehat{\mathbf{w}}$ is the minimizer of the problem (8), its ℓ_2-norm is no greater than any other feasible solution. Thus $\|\widehat{\mathbf{w}}\|_2 \leq \|\mathbf{w}_0^*\|_2$ and furthermore $\|\mathbf{w}_0^*\|_2$ is bounded by $\|\mathbf{b}_0^*\|_\infty/\sqrt{N}$ due to the definition of $\mathbf{w}_0^*$.

Part (b): The Lagrangian of (9) can be written as

$$\mathcal{L}(\boldsymbol{\alpha}, \gamma) = \frac{1}{2}\boldsymbol{\alpha}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\alpha} + \frac{1}{N}\mathbf{1}_N'\widehat{\boldsymbol{\Sigma}}\boldsymbol{\alpha} + \tau\,\|\boldsymbol{\alpha}\|_1 + \gamma\mathbf{1}_N'\boldsymbol{\alpha},$$

where γ is the Lagrangian multiplier for the constraint $\mathbf{1}_N'\boldsymbol{\alpha} = 0$. Consider the subgradient of $\mathcal{L}(\widehat{\boldsymbol{\alpha}}, \widehat{\gamma})$ with respect to α_i for any $i \in [N]$, where $(\widehat{\boldsymbol{\alpha}}, \widehat{\gamma})$ is the optimizer. Noting that $\widehat{\mathbf{A}}'\widehat{\mathbf{A}} = \widehat{\mathbf{A}}'\widehat{\boldsymbol{\Sigma}}$ due to the fact that $\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'$ is a projection matrix, the KKT conditions imply that

$$\left|\widehat{\boldsymbol{\alpha}}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}_{\cdot i} + \frac{1}{N}\mathbf{1}_N'\widehat{\boldsymbol{\Sigma}}_{\cdot i} + \widehat{\gamma}\right| = \left|(\widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}} + \frac{1}{N}\mathbf{1}_N)'\widehat{\boldsymbol{\Sigma}}_{\cdot i} + \widehat{\gamma}\right| = \left|\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}_{\cdot i} + \widehat{\gamma}\right| \leq \tau \text{ for all } i \in [N]$$

and furthermore

$$\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}_{\cdot i} + \widehat{\gamma} = \tau\text{sign}\,(\widehat{\alpha}_i) \text{ for all } \widehat{\alpha}_i \neq 0. \tag{42}$$

Suppose $\widehat{\boldsymbol{\alpha}} \notin \mathbb{S}^{\text{all}}$. Without loss of generality, let $\widehat{\alpha}_i > 0$ and $\widehat{\alpha}_j < 0$ for some $i, j \in \mathcal{G}_k$, $i \neq j$. (42) indicates $\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}_{\cdot i} + \widehat{\gamma} = \tau$ and $\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}_{\cdot j} + \widehat{\gamma} = -\tau$. Subtracting these two equations on both sides yields

$$\begin{aligned}2\tau &= \left|\widehat{\mathbf{w}}'(\widehat{\boldsymbol{\Sigma}}_{\cdot i} - \widehat{\boldsymbol{\Sigma}}_{\cdot j})\right| = \left|\widehat{\mathbf{w}}'[(\widehat{\boldsymbol{\Sigma}}^e_{\cdot i} + \widehat{\boldsymbol{\Sigma}}^*_{\cdot i}) - (\widehat{\boldsymbol{\Sigma}}^e_{\cdot j} + \widehat{\boldsymbol{\Sigma}}^*_{\cdot j})]\right| = \left|\widehat{\mathbf{w}}'(\widehat{\boldsymbol{\Sigma}}^e_{\cdot i} - \widehat{\boldsymbol{\Sigma}}^e_{\cdot j})\right| \\ &\leq \|\widehat{\boldsymbol{\Sigma}}^e_{\cdot i} - \widehat{\boldsymbol{\Sigma}}^e_{\cdot j}\|_2\,\|\widehat{\mathbf{w}}\|_2 \leq 2\|\widehat{\boldsymbol{\Sigma}}^e\|_{c2}\,\|\widehat{\mathbf{w}}\|_2 \leq 2\phi_e\,\|\mathbf{b}_0^*\|_\infty/\sqrt{N},\end{aligned} \tag{43}$$

where the third equality holds as $\widehat{\boldsymbol{\Sigma}}^*_{\cdot i} = \widehat{\boldsymbol{\Sigma}}^*_{\cdot j}$ for i and j in the same group k, and the last inequality by Part (a) which bounds $\|\widehat{\mathbf{w}}\|_2$. The above inequality (43) violates the presumption $\tau > \phi_e\,\|\mathbf{b}_0^*\|_\infty/\sqrt{N}$. We thus conclude $\widehat{\boldsymbol{\alpha}} \in \mathbb{S}^{\text{all}}$. ■

Proof of Lemma 3. For a generic vector $\boldsymbol{\delta} \in \tilde{\mathbb{S}}^{\text{all}}$, we have

$$\|\widehat{\mathbf{A}}\boldsymbol{\delta}\|_2 \geq \|\widehat{\mathbf{A}}^*\boldsymbol{\delta}\|_2 - \| \left(\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'\right) \widehat{\boldsymbol{\Sigma}}^e\boldsymbol{\delta}\|_2 \geq \|\widehat{\mathbf{A}}^*\boldsymbol{\delta}\|_2 - \|\widehat{\boldsymbol{\Sigma}}^e\boldsymbol{\delta}\|_2, \tag{44}$$

where the first inequality by the the triangle inequality, and the second follows because $\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'$ is a projection matrix. We will bound the two terms on the right-hand side.

To take advantage of the group structure to handle collinearity, we introduce a novel group-wise semi-norm and establish a corresponding version of compatibility condition. Let $d_k = \sum_{i\in\mathcal{G}_k} \delta_i$ for $k \leq K$ and $\mathbf{d} = (d_1, \ldots, d_K)'$. Define a group-wise ℓ_2 semi-norm $\|\cdot\|_{2\mathcal{G}} : \mathbb{R}^N \mapsto \mathbb{R}^+$ as

$$\|\boldsymbol{\delta}\|_{2\mathcal{G}} = \|\mathbf{d}\|_2 \,. \tag{45}$$

The definition of the semi-norm depends on the true group membership, which is infeasible in reality. We introduce this semi-norm only for theoretical development. In estimation we do not need to know the true group membership. This semi-norm $\|\boldsymbol{\delta}\|_{2\mathcal{G}}$ allows $\|\boldsymbol{\delta}\|_{2\mathcal{G}} = 0$ even if $\boldsymbol{\delta} \neq \mathbf{0}_N$, while it remains homogeneous, sub-additive, and non-negative— all other desirable properties of a norm. Moreover, if $\boldsymbol{\delta} \in \mathbb{S}^{\text{all}}$ it is obvious

$$\|\boldsymbol{\delta}\|_1 = \sum_{k\in[K]} \left| \sum_{i\in\mathcal{G}_k} \delta_i \right| = \sum_{k\in[K]} |d_k| \leq \sqrt{K}\, \|\boldsymbol{\delta}\|_{2\mathcal{G}} \tag{46}$$

by either the Cauchy-Schwarz or Jensen's inequality.

For any $\boldsymbol{\delta} \in \tilde{\mathbb{S}}^{\text{all}}$, we have

$$\begin{aligned} \|\widehat{\mathbf{A}}^*\boldsymbol{\delta}\|_2 &= \sqrt{N}\|\widehat{\mathbf{A}}^{\text{co}}\mathbf{d}\|_2 = \sqrt{N} \left\| \begin{pmatrix} \widehat{\mathbf{A}}^{\text{co}}\mathbf{d} \\ 0 \end{pmatrix} \right\|_2 = \sqrt{N}\|\tilde{\mathbf{A}}^{\text{co}}\mathbf{d}\|_2 \\ &\geq \sqrt{N\phi_A}\, \|\mathbf{d}\|_2 = \sqrt{N\phi_A}\, \|\boldsymbol{\delta}\|_{2\mathcal{G}} \geq \sqrt{\frac{N\phi_A}{K}}\, \|\boldsymbol{\delta}\|_1 \,, \end{aligned} \tag{47}$$

where the first equality follows by Lemma 2(a), the third equality by $\mathbf{1}_K'\mathbf{d} = \mathbf{1}_N'\boldsymbol{\delta} = 0$, and the last inequality by (46). We have found a lower bound for the first term on the right-hand side of (44). For the second term on the right-hand side of (44), we have

$$\|\widehat{\boldsymbol{\Sigma}}^e\boldsymbol{\delta}\|_2 \leq \max_i \|\widehat{\boldsymbol{\Sigma}}^e_{\cdot i}\|_2 \, \|\boldsymbol{\delta}\|_1 \leq \phi_e \, \|\boldsymbol{\delta}\|_1 \tag{48}$$

by (78) and the definition of ϕ_e.

Collecting (44) and (47)-(48), we have

$$\|\widehat{\mathbf{A}}\boldsymbol{\delta}\|_2 \geq \left(\sqrt{N\phi_A/K} - \phi_e\right) \|\boldsymbol{\delta}\|_1 \geq \frac{1}{2}\sqrt{N\phi_A/K}\, \|\boldsymbol{\delta}\|_1$$

under the presumption $\phi_e \leq \frac{1}{2}\sqrt{N\phi_A/K}$. The conclusion follows. ■

A.3 Proofs of the Results in Section 4

Proof of Lemma 4. Part (a). By the definition of ϕ_e and the triangle inequality,

$$\begin{aligned} \phi_e &= \|\widehat{\boldsymbol{\Sigma}}^e\|_{c2} \leq \|\boldsymbol{\Sigma}_0^e\|_{c2} + \|\boldsymbol{\Delta}^e\|_{c2} \\ &\leq C_{e0}\phi_{\max}(\boldsymbol{\Sigma}_0^e) + \sqrt{N}\left\|\boldsymbol{\Delta}^e\right\|_\infty = O_p(\sqrt{N}\phi_{NT}), \end{aligned} \tag{49}$$

where the second inequality and the last equality follow by Assumption 1(a).

Noting that $\widehat{\boldsymbol{\Sigma}}^{\text{co}} = \boldsymbol{\Sigma}_0^{\text{co}} + \boldsymbol{\Delta}^{\text{co}}$, and then w.p.a.1

$$\begin{aligned} \phi_{\min}(\widehat{\boldsymbol{\Sigma}}^{\text{co}}) &\geq \phi_{\min}(\boldsymbol{\Sigma}_0^{\text{co}}) - \left\|\boldsymbol{\Delta}^{\text{co}}\right\|_{\text{sp}} \geq \phi_{\min}(\boldsymbol{\Sigma}_0^{\text{co}}) - K\left\|\boldsymbol{\Delta}^{\text{co}}\right\|_\infty \\ &\geq \underline{c} - O_p(K(T/\log N)^{-1/2}) \geq \underline{c}/2 \end{aligned} \tag{50}$$

where the first inequality follows by Weyl inequality, the second inequality by the Gershgorin circle theorem, and the third inequality by Assumption 1(b). Similarly,

$$\phi_{\max}(\widehat{\boldsymbol{\Sigma}}^{\text{co}}) \leq \phi_{\max}(\boldsymbol{\Sigma}_0^{\text{co}}) + \left\|\boldsymbol{\Delta}^{\text{co}}\right\|_{\text{sp}} \leq 2\overline{c} \tag{51}$$

Suppose (50) and (51) occur. Given Assumption 2(b) about the rate of $\underline{r}$, (22) implies

$$\phi_A^{-1} \leq 8\underline{r}^{-1}\underline{c}^{-2} + \frac{4}{K}\frac{\overline{c}}{\underline{c}} = O_p\left(K\right)$$

and (18) implies

$$\begin{aligned} \|\mathbf{b}_0^*\|_\infty &\leq \left\|(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K\right\|_\infty / \left[\underline{r}\cdot\mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K\right] \leq \left(\mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-2}\mathbf{1}_K\right)^{1/2} / \left[\underline{r}\cdot\mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K\right] \\ &\leq \underline{r}^{-1}\phi_{\min}^{-1}\left(\widehat{\boldsymbol{\Sigma}}^{\text{co}}\right)\left(\mathbf{1}_K'(\widehat{\boldsymbol{\Sigma}}^{\text{co}})^{-1}\mathbf{1}_K\right)^{-1/2} \leq \underline{r}^{-1}K^{-1/2}\phi_{\max}^{1/2}\left(\widehat{\boldsymbol{\Sigma}}^{\text{co}}\right)/\phi_{\min}\left(\widehat{\boldsymbol{\Sigma}}^{\text{co}}\right) \\ &\leq 4\overline{c}^{1/2}/(\underline{rc}K^{1/2}) = O_p\left(\sqrt{K}\right). \end{aligned} \tag{52}$$

Part (b): The order of $\|\mathbf{b}_0^*\|_\infty$ in the above inequality gives $\phi_e\left\|\mathbf{b}_0^*\right\|_\infty/\sqrt{N} = O_p(K^{1/2}\phi_{NT})$. On the other hand, given the choice τ in (23), there exists a finite C_τ sufficient large such that $\tau > \phi_e\left\|\mathbf{b}_0^*\right\|_\infty/\sqrt{N}$. Consequently, the event $\left\{\phi_e\left\|\mathbf{b}_0^*\right\|_\infty/\sqrt{N} < \tau\right\}$ occurs w.p.a.1. ■

Proof of Theorem 2. When the sample size is sufficiently large we have $\left\{\phi_e\left\|\mathbf{b}_0^*\right\|_\infty/\sqrt{N} < \tau\right\}$ w.p.a.1 by Lemma 4(b). We can then construct the desirable $\widehat{\boldsymbol{\alpha}}^*$ according to (24). Since $\widehat{\boldsymbol{\alpha}}$ is the solution to (31),

$$\frac{1}{2}\widehat{\boldsymbol{\alpha}}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}} + \frac{1}{N}\mathbf{1}_N'\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}} + \tau\left\|\boldsymbol{\alpha}\right\|_1 \leq \frac{1}{2}\widehat{\boldsymbol{\alpha}}^{*\prime}\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}}^* + \frac{1}{N}\mathbf{1}_N'\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{\alpha}}^* + \tau\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1.$$

Define $\boldsymbol{\psi} = \widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*$. Rearranging the above inequality yields

$$\boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\psi} + 2\tau \left\|\widehat{\boldsymbol{\alpha}}\right\|_1 \leq -2\boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}}^* + \frac{\mathbf{1}_N}{N}) + 2\tau \left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1 . \tag{53}$$

Notice that

$$\begin{aligned}
&\boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}\left(\widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}}^* + \frac{\mathbf{1}_N}{N}\right) \\
&= \boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{A}}^*\widehat{\boldsymbol{\alpha}}^* + \frac{\mathbf{1}_N}{N}) + \boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{A}} - \widehat{\mathbf{A}}^*)\widehat{\boldsymbol{\alpha}}^* = \boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}\mathbf{w}_0^* + \boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}'\left(\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'\right)\widehat{\boldsymbol{\Sigma}}^e\widehat{\boldsymbol{\alpha}}^* \\
&= (\boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}^*\mathbf{w}_0^* + \boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}^e\mathbf{w}_0^*) + \boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\boldsymbol{\Sigma}}^e\widehat{\boldsymbol{\alpha}}^* = (-\gamma_0^*\boldsymbol{\psi}'\mathbf{1}_N + \boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}^e\mathbf{w}_0^*) + \boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\boldsymbol{\Sigma}}^e\widehat{\boldsymbol{\alpha}}^* \\
&= \boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}^e\mathbf{w}_0^* + \boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\boldsymbol{\Sigma}}^e\widehat{\boldsymbol{\alpha}}^*,
\end{aligned} \tag{54}$$

where the fourth equality follows by the fact that $\widehat{\boldsymbol{\Sigma}}^*\mathbf{w}_0^* = -\gamma_0^*\mathbf{1}_N$ implied by the KKT conditions in (5) with $\tau = 0$, and the last equality by $\boldsymbol{\psi}'\mathbf{1}_N = (\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*)'\mathbf{1}_N = 0$ as the dual problems entail both $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\alpha}}^*$ sum up to 0. Plugging (54) into (53) to bound the right-hand side of (53), we have

$$\boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\psi} + 2\tau\left\|\widehat{\boldsymbol{\alpha}}\right\|_1 \leq 2\left|\boldsymbol{\psi}'\widehat{\boldsymbol{\Sigma}}^e\mathbf{w}_0^* + \boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\boldsymbol{\Sigma}}^e\widehat{\boldsymbol{\alpha}}^*\right| + 2\tau\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1 \leq 2\left\|\boldsymbol{\psi}\right\|_1(\zeta_1 + \zeta_2) + 2\tau\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1 , \tag{55}$$

where $\zeta_1 = \left\|\widehat{\boldsymbol{\Sigma}}^e\mathbf{w}_0^*\right\|_\infty$ and $\zeta_2 = \left\|\widehat{\mathbf{A}}'\widehat{\boldsymbol{\Sigma}}^e\widehat{\boldsymbol{\alpha}}^*\right\|_\infty$.

Now we bound ζ_1 and ζ_2 in turn. By (41), (49), Lemma 4(a) and the choice of τ in (23), we have

$$\zeta_1 \leq \left\|\widehat{\boldsymbol{\Sigma}}^e\right\|_{c2}\left\|\mathbf{w}_0^*\right\|_2 = \phi_e C_b/\sqrt{N} = O_p\left(K^{1/2}\phi_{NT}\right)$$

For ζ_2, we have by (77),

$$\zeta_2 \leq \left\|\widehat{\mathbf{A}}\right\|_{c2}\left\|\widehat{\boldsymbol{\Sigma}}^e\right\|_{c2}\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1 = \left\|\widehat{\mathbf{A}}'\right\|_{c2}\cdot\phi_e\cdot\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1 . \tag{56}$$

Noting that $\left\|\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'\right\|_{\text{sp}} = 1$ and by the triangle inequality, we have

$$\begin{aligned}
\left\|\widehat{\mathbf{A}}'\right\|_{c2} &\leq \left\|\widehat{\boldsymbol{\Sigma}}\right\|_{c2} \leq \left\|\widehat{\boldsymbol{\Sigma}}^*\right\|_{c2} + \left\|\widehat{\boldsymbol{\Sigma}}^e\right\|_{c2} + \phi_e \\
&\leq \sqrt{N}\left(\phi_{\max}\left(\boldsymbol{\Sigma}_0^{\text{co}}\right) + \left\|\boldsymbol{\Delta}^{\text{co}}\right\|_\infty\right) + O_p(\sqrt{N}\phi_{NT}) = O_p(\sqrt{N}),
\end{aligned}$$

where the third inequality follows from Assumption 1(b) and the order of ϕ_e by Lemma 4(a). Noting that $\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1 = \left\|\mathbf{a}_0^*\right\|_1 \leq \left(\left\|\mathbf{b}_0^*\right\|_\infty + 1\right)\sqrt{K/\phi_A}/N$ by Lemma 2, we continue (56) to obtain

$$\zeta_2 = O_p(\sqrt{N})O_p(\sqrt{N}\phi_{NT})\left\|\mathbf{b}_0^*\right\|_\infty\sqrt{K/\phi_A}N^{-1} = O_p(K^{1/2}\phi_{NT}\sqrt{K/\phi_A})$$

where the second equality follows by Assumption 2(a) and $\left\|\mathbf{b}_0^*\right\|_\infty = O_p\left(K^{1/2}\right)$ by Lemma 4(a), and the last equality by (23). We thus obtain $\zeta_1 + \zeta_2 = o_p(K^{1/2}\phi_{NT}\sqrt{K/\phi_A})$.

Now, suppose that the sample size is sufficiently large, there exists a finite C_τ such that $\zeta_1 + \zeta_2 \leq$

$\tau\sqrt{K/\phi_A}/2$. We push (55) further to attain

$$\boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\psi} + 2\tau\left\|\widehat{\boldsymbol{\alpha}}\right\|_1 \leq \tau\sqrt{K/\phi_A}\left\|\boldsymbol{\psi}\right\|_1 + 2\tau\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1.$$

Then

$$\boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\psi} \leq \tau\sqrt{K/\phi_A}\left\|\boldsymbol{\psi}\right\|_1 + 2\tau\left(\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1 - \left\|\widehat{\boldsymbol{\alpha}}\right\|_1\right) \leq \tau\left(\sqrt{K/\phi_A}+2\right)\left\|\boldsymbol{\psi}\right\|_1$$

where the last inequality follows by the triangle inequality: $\left\|\widehat{\boldsymbol{\alpha}}^*\right\|_1 - \left\|\widehat{\boldsymbol{\alpha}}\right\|_1 \leq \left\|\boldsymbol{\psi}\right\|_1$. Adding $\tau\sqrt{K/\phi_A}\left\|\boldsymbol{\psi}\right\|_1$ to both sides of the above inequality yields

$$\boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\psi} + \tau\sqrt{K/\phi_A}\left\|\boldsymbol{\psi}\right\|_1 \leq 2\tau\left(\sqrt{K/\phi_A}+1\right)\left\|\boldsymbol{\psi}\right\|_1 \leq 4\tau\sqrt{K/\phi_A}\left\|\boldsymbol{\psi}\right\|_1 \tag{57}$$

where the last inequality follows the fact $K/\phi_A \geq 1$ in view of $K \geq 1$ and $\phi_A \leq 1$ by definition.

By Lemma 4, we have

$$\sqrt{\frac{K/\phi_A}{N}}\phi_e = \sqrt{\frac{K/\phi_A}{N}}O_p\left(\sqrt{N}\phi_{NT}\right) = O_p\left(\sqrt{K/\phi_A}\phi_{NT}\right) = O_p\left(K\phi_{NT}\right) = o_p\left(1\right).$$

where the third equality holds because $\sqrt{K/\phi_A} = O_p\left(K\right)$ by Lemma 4(a) and the last holds by Assumption 2(a). This implies that the condition $\phi_e \leq \frac{1}{2}\sqrt{N\phi_A/K}$ in Lemma 3 is satisfied w.p.a.1. Moreover, $\boldsymbol{\psi} \in \tilde{\mathbb{S}}^{\text{all}}$ by construction of $\widehat{\boldsymbol{\alpha}}^*$ in (24). We hence invoke Lemma 3 to continue (57):

$$4\tau\sqrt{K/\phi_A}\left\|\boldsymbol{\psi}\right\|_1 \leq 8\tau\frac{K/\phi_A}{\sqrt{N}}\|\widehat{\mathbf{A}}\boldsymbol{\psi}\|_2 \leq \frac{1}{2}\boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\psi} + 32\tau^2\frac{(K/\phi_A)^2}{N} \tag{58}$$

where the last inequality follows by $8ab \leq \frac{1}{2}a^2 + 32b^2$. Combining (57) and (58), we have

$$\frac{1}{2}\boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\psi} + \tau\sqrt{K/\phi_A}\left\|\boldsymbol{\psi}\right\|_1 \leq 32\tau^2\frac{(K/\phi_A)^2}{N}. \tag{59}$$

The above equality immediately implies

$$\left\|\boldsymbol{\psi}\right\|_1 = \left\|\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*\right\|_1 \leq 32\tau\frac{(K/\phi_A)^{3/2}}{N} = O_p\left(\frac{(K/\phi_A)^{3/2}\tau}{N}\right) = O_p\left(\frac{K^3\tau}{N}\right)$$

and

$$\sqrt{\boldsymbol{\psi}'\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\boldsymbol{\psi}} = \left\|\widehat{\mathbf{A}}\left(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*\right)\right\|_2 \leq 8\tau\frac{K/\phi_A}{\sqrt{N}} = O_p\left(\frac{(K/\phi_A)\tau}{\sqrt{N}}\right) = O_p\left(\frac{K^2\tau}{\sqrt{N}}\right),$$

given $K/\phi_A = O_p\left(K^2\right)$ by Lemma 4(a). This completes the proof of the theorem. ■

Proof of Corollary 1. Recall that $\widehat{\mathbf{A}}^* = (\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N')\widehat{\boldsymbol{\Sigma}}^*$ and $\widehat{\mathbf{A}} = (\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N')\widehat{\boldsymbol{\Sigma}}$. Let $\widehat{\mathbf{A}}^e := (\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N')\widehat{\boldsymbol{\Sigma}}^e$. Then we have

$$\widehat{\mathbf{w}} - \mathbf{w}_0^* = \left(\widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}} + \frac{\mathbf{1}_N}{N}\right) - \left(\widehat{\mathbf{A}}^*\widehat{\boldsymbol{\alpha}}^* + \frac{\mathbf{1}_N}{N}\right) = \widehat{\mathbf{A}}\widehat{\boldsymbol{\alpha}} - \left(\widehat{\mathbf{A}} - \widehat{\mathbf{A}}^e\right)\widehat{\boldsymbol{\alpha}}^* = \widehat{\mathbf{A}}\left(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*\right) + \widehat{\mathbf{A}}^e\widehat{\boldsymbol{\alpha}}^*. \tag{60}$$

For the first term in (60)

$$\|\widehat{\mathbf{A}}\left(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^{*}\right)\|_{2}=O_{p}\left(N^{-1/2}K^{2}\tau\right) \tag{61}$$

is already given by Theorem 2.

For the second term in (60), we have $\|\widehat{\mathbf{A}}^{e}\widehat{\boldsymbol{\alpha}}^{*}\|_{2}\leq\|\widehat{\boldsymbol{\Sigma}}^{e}\widehat{\boldsymbol{\alpha}}^{*}\|_{2}\leq\|\boldsymbol{\Sigma}_{0}^{e}\widehat{\boldsymbol{\alpha}}^{*}\|_{2}+\|\boldsymbol{\Delta}^{e}\widehat{\boldsymbol{\alpha}}^{*}\|_{2}:=I_{1}+I_{2}$ by the triangle inequality. Notice that

$$\begin{aligned}
I_{1} &\leq \phi_{\max}\left(\boldsymbol{\Sigma}_{0}^{e}\right)\|\widehat{\boldsymbol{\alpha}}^{*}\|_{2}\leq\phi_{\max}\left(\boldsymbol{\Sigma}_{0}^{e}\right)\|\mathbf{a}_{0}^{*}\|_{2}\\
&\leq \phi_{\max}\left(\boldsymbol{\Sigma}_{0}^{e}\right)\frac{1}{N\sqrt{\phi_{A}}}\left(\|\mathbf{b}_{0}^{*}\|_{\infty}+1\right)=O_{p}\left(\frac{K^{1/2}\phi_{NT}}{\sqrt{N\phi_{A}}}\right),
\end{aligned} \tag{62}$$

where the second inequality follows as $\widehat{\boldsymbol{\alpha}}^{*}\in\tilde{\mathbb{S}}^{\text{all}}\subset\mathbb{S}^{\text{all}}$ by construction, the third inequality by (39), and last equality holds by Assumption 1(a) and Lemma 4(a). Moreover,

$$\begin{aligned}
I_{2} &\leq \|\boldsymbol{\Delta}^{e}\|_{c2}\|\widehat{\boldsymbol{\alpha}}^{*}\|_{1}\leq\sqrt{N}\|\boldsymbol{\Delta}^{e}\|_{\infty}\|\widehat{\boldsymbol{\alpha}}^{*}\|_{1}\\
&= \sqrt{N}\|\boldsymbol{\Delta}^{e}\|_{\infty}\|\mathbf{a}_{0}^{*}\|_{1}\leq\sqrt{N}\|\boldsymbol{\Delta}^{e}\|_{\infty}O_{p}\left(N^{-1}\sqrt{K/\phi_{A}}\left(\|\mathbf{b}_{0}^{*}\|_{\infty}+1\right)\right)\\
&= O_{p}\left(N^{-1/2}K^{1/2}\phi_{NT}\sqrt{K/\phi_{A}}\right),
\end{aligned} \tag{63}$$

where the first inequality follows by (78), the first equality holds by the fact that $\|\widehat{\boldsymbol{\alpha}}^{*}\|_{1}=\|\boldsymbol{\alpha}_{0}^{*}\|_{1}=\|\mathbf{a}_{0}^{*}\|_{1}$ as in (25), the third inequality by (40), and the last equality holds by Assumption 1(a) and Lemma 4(a).

Collecting (61), (62) and (63), we attain

$$\|\widehat{\mathbf{w}}-\mathbf{w}_{0}^{*}\|_{2}=N^{-1/2}O_{p}\left(K^{2}\tau+\sqrt{K/\phi_{A}}\phi_{NT}+K^{1/2}\phi_{NT}\sqrt{K/\phi_{A}}\right)=O_{p}\left(N^{-1/2}\tau K^{2}\right)=o_{p}\left(N^{-1/2}\right),$$

where and the last equality holds by Assumption 2(a). In addition, by the Cauchy-Schwarz inequality, we have $\|\widehat{\mathbf{w}}-\mathbf{w}_{0}^{*}\|_{1}\leq\sqrt{N}\|\widehat{\mathbf{w}}-\mathbf{w}_{0}^{*}\|_{2}=O_{p}\left(K^{2}\tau\right)=o_{p}\left(1\right).$ ■

Proof of Theorem 3. Part (a). Denote $\boldsymbol{\psi}_{w}=\widehat{\mathbf{w}}-\mathbf{w}_{0}^{*}$. We first show the in-sample oracle inequality. Decompose

$$\begin{aligned}
& \widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}}-\mathbf{w}_{0}^{**\prime}\widehat{\boldsymbol{\Sigma}}^{*}\mathbf{w}_{0}^{*}\\
= & \left(\mathbf{w}_{0}^{**\prime}\widehat{\boldsymbol{\Sigma}}\mathbf{w}_{0}^{*}+2\boldsymbol{\psi}_{w}'\widehat{\boldsymbol{\Sigma}}\mathbf{w}_{0}^{*}+\boldsymbol{\psi}_{w}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\psi}_{w}\right)-\mathbf{w}_{0}^{**\prime}\widehat{\boldsymbol{\Sigma}}^{*}\mathbf{w}_{0}^{*}=\mathbf{w}_{0}^{**\prime}\widehat{\boldsymbol{\Sigma}}^{e}\mathbf{w}_{0}^{*}+2\boldsymbol{\psi}_{w}'\widehat{\boldsymbol{\Sigma}}\mathbf{w}_{0}^{*}+\boldsymbol{\psi}_{w}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\psi}_{w}\\
= & \mathbf{w}_{0}^{**\prime}\widehat{\boldsymbol{\Sigma}}^{e}\mathbf{w}_{0}^{*}+2\boldsymbol{\psi}_{w}'(\widehat{\boldsymbol{\Sigma}}^{*}+\boldsymbol{\Delta}^{e})\mathbf{w}_{0}^{*}+2\boldsymbol{\psi}_{w}'\boldsymbol{\Sigma}_{0}^{e}\mathbf{w}_{0}^{*}+\boldsymbol{\psi}_{w}'(\widehat{\boldsymbol{\Sigma}}^{*}+\boldsymbol{\Delta}^{e})\boldsymbol{\psi}_{w}+\boldsymbol{\psi}_{w}'\boldsymbol{\Sigma}_{0}^{e}\boldsymbol{\psi}_{w}\\
= & :II_{1}+2II_{2}+2II_{3}+II_{4}+II_{5}.
\end{aligned}$$

We bound II_1 by

$$\begin{aligned}
|II_1| &\leq \phi_{\max}(\widehat{\mathbf{\Sigma}}^e)\left\|\mathbf{w}_0^*\right\|_2^2 \leq \left(\phi_{\max}\left(\mathbf{\Sigma}_0^e\right) + \phi_{\max}\left(\mathbf{\Delta}^e\right)\right)\left\|\mathbf{w}_0^*\right\|_2^2 \\
&\leq \left(\phi_{\max}\left(\mathbf{\Sigma}_0^e\right) + N\|\mathbf{\Delta}^e\|_\infty\right)\left\|\mathbf{w}_0^*\right\|_2^2 \\
&\leq \left(O_p(\sqrt{N}\phi_{NT}) + NO_p((T/\log N)^{-1/2})\right)\frac{\|\mathbf{b}_0^*\|_\infty^2}{N} = O_p(K(T/\log N)^{-1/2}),
\end{aligned}$$

where the third inequality holds by the Gershgorin circle theorem, and the fourth inequality by Assumption 1 and Lemma 4(a). The second term II_2 is bounded by

$$\begin{aligned}
|II_2| &\leq \|\widehat{\mathbf{\Sigma}}^* + \mathbf{\Delta}^e\|_\infty \left\|\boldsymbol{\psi}_w\right\|_1 \left\|\mathbf{w}_0^*\right\|_1 \leq \left(\|\widehat{\mathbf{\Sigma}}^*\|_\infty + \|\mathbf{\Delta}^e\|_\infty\right)\left\|\boldsymbol{\psi}_w\right\|_1 \left\|\mathbf{w}_0^*\right\|_1 \\
&\leq \left(\|\widehat{\mathbf{\Sigma}}^{\mathrm{co}}\|_\infty + \|\mathbf{\Delta}^e\|_\infty\right)\left\|\boldsymbol{\psi}_w\right\|_1 \sqrt{N}\left\|\mathbf{w}_0^*\right\|_2 \\
&= O_p\left(\bar{c} + (T/\log N)^{-1/2}\right) O_p\left(\tau K^2\right)\|\mathbf{b}_0^*\|_\infty = O_p\left(\tau K^{5/2}\right),
\end{aligned}$$

where the first inequality follows by (75), the third inequality by the Cauchy-Schwarz inequality, the first equality holds by Assumptions 1 and Corollary 1, and the last equality holds by Assumption 2(b) and and Lemma 4(a). For II_3, we have

$$\begin{aligned}
|II_3| &\leq \phi_{\max}\left(\mathbf{\Sigma}_0^e\right)\left\|\boldsymbol{\psi}_w\right\|_2\left\|\mathbf{w}_0^*\right\|_2 \\
&= O_p(\sqrt{N}\phi_{NT})O_p\left(N^{-1/2}\tau K^2\right)\|\mathbf{b}_0^*\|_\infty N^{-1/2} = O_p\left(N^{-1/2}\phi_{NT}\tau K^{5/2}\right)
\end{aligned}$$

by (76), Assumptions 1, and Corollary 1. Similarly,

$$\begin{aligned}
|II_4| &\leq \|\widehat{\mathbf{\Sigma}}^* + \mathbf{\Delta}^e\|_\infty \left\|\boldsymbol{\psi}_w\right\|_1^2 = O_p(\bar{c} + (T/\log N)^{-1/2})O_p\left(\tau^2 K^4\right) = O_p\left(\tau^2 K^4\right), \text{ and} \\
|II_5| &\leq \phi_{\max}\left(\mathbf{\Sigma}_0^e\right)\left\|\boldsymbol{\psi}_w\right\|_2^2 = O_p(\sqrt{N}\phi_{NT})\left(N^{-1}\tau^2 K^4\right) = O_p\left(N^{-1/2}\phi_{NT}\tau^2 K^4\right).
\end{aligned}$$

Collecting all these five terms, and notice that $\tau K^{5/2}$ is the dominating order, we have

$$\left|\widehat{\mathbf{w}}'\widehat{\mathbf{\Sigma}}\widehat{\mathbf{w}} - \mathbf{w}_0^{**\prime}\widehat{\mathbf{\Sigma}}^*\mathbf{w}_0^*\right| \leq O_p\left(\tau K^{5/2}\right),$$

and the in-sample oracle inequality holds.

Part (b). The same argument goes through if we replace $\widehat{\mathbf{\Sigma}}$ with $\widehat{\mathbf{\Sigma}}^{\mathrm{new}}$, and replace $\widehat{\mathbf{\Sigma}}^*$ with $\widehat{\mathbf{\Sigma}}^{*\mathrm{new}}$, because in the above analysis of the in-sample oracle inequality we always bound the various quantities by separating the norms of vectors and the square matrices. We conclude the out-of-sample oracle inequality.

Part (d). This proof involves two steps. (i) Establish the closeness between $\widehat{\mathbf{w}}'\widehat{\mathbf{\Sigma}}^{\mathrm{new}}\widehat{\mathbf{w}}$ and $\widehat{\mathbf{w}}'\widehat{\mathbf{\Sigma}}\widehat{\mathbf{w}}$ as shown in (68) below; (ii) Establish the closeness between $\widehat{\mathbf{w}}'\widehat{\mathbf{\Sigma}}\widehat{\mathbf{w}}$ and $Q(\mathbf{\Sigma}_0)$ as shown in (73) below.

Let $Q(\mathbf{S})$ be the minimum of $\min_{\mathbf{w}\in\mathbb{R}^N, \mathbf{w}'\mathbf{1}_N=1} \mathbf{w}'\mathbf{S}\mathbf{w}$ for a generic positive semi-definite matrix

S. Obviously $\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}^{\text{new}}\widehat{\mathbf{w}} \geq \widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} = Q(\widehat{\boldsymbol{\Sigma}})$. On the other hand,

$$\widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} = \widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}^{\text{new}}\widehat{\mathbf{w}} + \widehat{\mathbf{w}}'\left(\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^{\text{new}}\right)\widehat{\mathbf{w}} \geq \widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}^{\text{new}}\widehat{\mathbf{w}} - \|\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^{\text{new}}\|_{\text{sp}}\|\widehat{\mathbf{w}}\|_2^2 \tag{64}$$

by the triangle inequality and (76). We focus on the term $\|\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^{\text{new}}\|_{\text{sp}}\|\widehat{\mathbf{w}}\|_2^2$.

For the first factor, notice

$$\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^{\text{new}} = \widehat{\boldsymbol{\Sigma}}^* - \widehat{\boldsymbol{\Sigma}}^{*,\text{new}} + \widehat{\boldsymbol{\Sigma}}^e - \widehat{\boldsymbol{\Sigma}}^{e,\text{new}} = \left(\widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}_0^*\right) - \left(\widehat{\boldsymbol{\Sigma}}^{*,\text{new}} - \boldsymbol{\Sigma}_0^*\right) + \boldsymbol{\Delta}^e - \boldsymbol{\Delta}^{e,\text{new}}.$$

Under Assumption 1(b), $\left\|\boldsymbol{\Sigma}_0^* - \widehat{\boldsymbol{\Sigma}}_0^*\right\|_\infty = \left\|\boldsymbol{\Sigma}_0^{\text{co}} - \widehat{\boldsymbol{\Sigma}}^{\text{co}}\right\|_\infty = \|\Delta^{\text{co}}\|_\infty = O_p((T/\log N)^{-1/2})$ and therefore by the Gershgorin circle theorem the spectral norm is bounded by

$$\|\boldsymbol{\Sigma}_0^* - \widehat{\boldsymbol{\Sigma}}^*\|_{\text{sp}} \leq N\|\Delta^{\text{co}}\|_\infty = O_p\left(N(T/\log N)^{-1/2}\right). \tag{65}$$

By the Gershgorin circle theorem, $\|\boldsymbol{\Delta}^e\|_{\text{sp}} \leq \phi_e = O_p(\sqrt{N}\phi_{NT})$ by Lemma 4(a). Since the new testing data comes from the same data generating process as that of the training data, the same stochastic bounds is applicable to the terms involving the new data, and then

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^{\text{new}}\|_{\text{sp}} &\leq \|\widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}_0^*\|_{\text{sp}} + \|\widehat{\boldsymbol{\Sigma}}^{*,\text{new}} - \boldsymbol{\Sigma}_0^*\|_{\text{sp}} + \|\boldsymbol{\Delta}^e\|_{\text{sp}} + \|\boldsymbol{\Delta}^{e,\text{new}}\|_{\text{sp}} \\ &= O_p\left(N(T/\log N)^{-1/2}\right) + O_p(\sqrt{N}\phi_{NT}) = O_p(N\phi_{NT}). \end{aligned} \tag{66}$$

The second factor is bounded by

$$\|\widehat{\mathbf{w}}\|_2^2 \leq \|\mathbf{b}_0^*\|_\infty^2/N = O_p(K/N) \tag{67}$$

according to Theorem 1(a) and Lemma 4(a). Collecting (64), (67) and (66), we have

$$0 \leq \widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}^{\text{new}}\widehat{\mathbf{w}} - \widehat{\mathbf{w}}'\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{w}} \leq \|\widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}^{\text{new}}\|_{\text{sp}}\|\widehat{\mathbf{w}}\|_2^2 = O_p(N\phi_{NT})O_p(K/N) = O_p(K\phi_{NT}). \tag{68}$$

Next, consider the population matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_0^*$. Because $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0^* = \boldsymbol{\Sigma}_0^e$ is positive semi-definite,

$$Q(\boldsymbol{\Sigma}_0) \geq Q(\boldsymbol{\Sigma}_0^*). \tag{69}$$

Since $\text{rank}(\boldsymbol{\Sigma}_0^*) = K \ll N$, the solution to $\min_{\mathbf{w}\in\mathbb{R}^N, \mathbf{w}'\mathbf{1}_N=1} \mathbf{w}'\boldsymbol{\Sigma}_0^*\mathbf{w}$ is not unique but all the solutions give the same minimum $Q(\boldsymbol{\Sigma}_0^*)$. Thus in order to evaluate $Q(\boldsymbol{\Sigma}_0^*)$ we can simply use the within-group equal weight solution $\mathbf{w}_0^\sharp$ corresponding to

$$\min_{(\mathbf{w},\gamma)\in\mathbb{R}^{N+1}} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{subject to} \quad \mathbf{w}'\mathbf{1}_N = 1, \text{ and } \boldsymbol{\Sigma}_0^*\mathbf{w} + \gamma = 0.$$

The only difference between $\mathbf{w}_0^\sharp$ and $\mathbf{w}_0^*$ is that the former is associated with the population $\boldsymbol{\Sigma}_0^*$ and the latter associated with the sample $\widehat{\boldsymbol{\Sigma}}^*$. Parallel to (17), (18) and (52), we can write $\mathbf{w}_0^\sharp =$

$\left(\frac{b^\sharp_{01}}{N}\cdot \mathbf{1}'_{N_1},\cdots,\frac{b^\sharp_{0K}}{N}\cdot \mathbf{1}'_{N_K}\right)'$ where $b^\sharp_{0k} = \frac{[(\mathbf{\Sigma}_0^{\text{co}})^{-1}\mathbf{1}_K]_k}{r_k\cdot \mathbf{1}'_K(\mathbf{\Sigma}_0^{\text{co}})^{-1}\mathbf{1}_K}$, and it is bounded by

$$\begin{aligned}\left\|\mathbf{b}_0^\sharp\right\|_\infty &\le \left\|(\mathbf{\Sigma}_0^{\text{co}})^{-1}\mathbf{1}_K\right\|_\infty / \left[\underline{r}\cdot\mathbf{1}'_K(\mathbf{\Sigma}_0^{\text{co}})^{-1}\mathbf{1}_K\right] \le \underline{r}^{-1}K^{-1/2}\phi_{\max}^{1/2}\left(\mathbf{\Sigma}_0^{\text{co}}\right)/\phi_{\min}\left(\mathbf{\Sigma}_0^{\text{co}}\right)\\ &\le \bar{c}^{1/2}/(\underline{rc}K^{1/2}) = O\left(\sqrt{K}\right)\end{aligned}$$

under Assumption 1(b) and furthermore

$$\|\mathbf{w}_0^\sharp\|_2^2 \le N\|\mathbf{w}_0^\sharp\|_\infty^2 = N(\|\mathbf{b}_0^\sharp\|_\infty/N)^2 = O\left(K/N\right). \tag{70}$$

We continue (69):

$$\begin{aligned}Q(\mathbf{\Sigma}_0^*) = \mathbf{w}_0^{\sharp\prime}\widehat{\mathbf{\Sigma}}^*\mathbf{w}_0^\sharp + \mathbf{w}_0^{\sharp\prime}\left(\mathbf{\Sigma}_0^* - \widehat{\mathbf{\Sigma}}^*\right)\mathbf{w}_0^\sharp &\ge \mathbf{w}_0^{*\prime}\widehat{\Sigma}^*\mathbf{w}_0^* + \mathbf{w}_0^{\sharp\prime}\left(\mathbf{\Sigma}_0^* - \widehat{\mathbf{\Sigma}}^*\right)\mathbf{w}_0^\sharp\\ &\ge \mathbf{w}_0^{*\prime}\widehat{\Sigma}^*\mathbf{w}_0^* - \|\mathbf{\Sigma}_0^* - \widehat{\mathbf{\Sigma}}^*\|_{\text{sp}}\|\mathbf{w}_0^\sharp\|_2^2,\end{aligned} \tag{71}$$

where the first inequality follows as $\mathbf{w}_0^*$ is the optimizer associated with $\widehat{\Sigma}^*$, and the second inequality is derived by the same reasoning as (64). It implies

$$\mathbf{w}_0^{*\prime}\widehat{\Sigma}^*\mathbf{w}_0^* \le Q(\mathbf{\Sigma}_0) + \|\mathbf{\Sigma}_0^* - \widehat{\mathbf{\Sigma}}^*\|_{\text{sp}}\|\mathbf{w}_0^\sharp\|_2^2 \le Q(\mathbf{\Sigma}_0) + O_p\left(K(T/\log N)^{-1/2}\right) \tag{72}$$

in view of (66) and (70). Combine Part (a) and (72):

$$\widehat{\mathbf{w}}'\widehat{\mathbf{\Sigma}}\widehat{\mathbf{w}} \le Q(\mathbf{\Sigma}_0) + O_p\left(\tau K^{5/2}\right). \tag{73}$$

In conjunction with (68) and notice $K\phi_{NT}$ is of smaller order than $\tau K^{5/2}$, the conclusion follows.■

A.4 Elementary Inequalities on Matrix Norms

We collect some elementary inequalities used in the proofs.

Lemma 5. *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors, and* $\mathbf{A}$ *and* $\mathbf{B}$ *be two matrices of compatible dimensions. Then we have*

$$\|\mathbf{Ab}\|_\infty \le \|\mathbf{A}\|_\infty\|\mathbf{b}\|_1 \tag{74}$$

$$\left|\mathbf{a}'\mathbf{Ab}\right| \le \|\mathbf{A}\|_\infty\|\mathbf{a}\|_1\|\mathbf{b}\|_1 \tag{75}$$

$$\left|\mathbf{a}'\mathbf{Ab}\right| \le \|\mathbf{A}\|_{\text{sp}}\|\mathbf{a}\|_2\|\mathbf{b}\|_2 \tag{76}$$

$$\|\mathbf{ABb}\|_\infty \le \left\|\mathbf{A}'\right\|_{c2}\|\mathbf{B}\|_{c2}\|\mathbf{b}\|_1 \tag{77}$$

If $\mathbf{S}$ *is a symmetric matrix,*

$$\|\mathbf{Sb}\|_2 \le \|\mathbf{S}\|_{c2}\|\mathbf{b}\|_1. \tag{78}$$

If $\boldsymbol{\Sigma}$ is positive definite,

$$\left(\boldsymbol{\Sigma}+\mathbf{a}\mathbf{a}'\right)^{-1}=\boldsymbol{\Sigma}^{-1}-\frac{\boldsymbol{\Sigma}^{-1}\mathbf{a}\mathbf{a}'\boldsymbol{\Sigma}^{-1}}{1+\mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a}} \tag{79}$$

$$\left(\boldsymbol{\Sigma}-\mathbf{a}\mathbf{a}'\right)^{-1}=\boldsymbol{\Sigma}^{-1}+\frac{\boldsymbol{\Sigma}^{-1}\mathbf{a}\mathbf{a}'\boldsymbol{\Sigma}^{-1}}{1-\mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{a}}. \tag{80}$$

Proof. The first inequality follows because

$$\left\|\mathbf{A}\mathbf{b}\right\|_{\infty}\leq\max_{i}\left|\mathbf{A}_{i\cdot}\mathbf{b}\right|\leq\max_{i}\left\|\mathbf{A}_{i\cdot}\right\|_{\infty}\left\|\mathbf{b}\right\|_{1}=\left\|\mathbf{A}\right\|_{\infty}\left\|\mathbf{b}\right\|_{1}.$$

It implies the second inequality $\left|\mathbf{a}'\mathbf{A}\mathbf{b}\right|\leq\left\|\mathbf{a}\right\|_{1}\left\|\mathbf{A}\mathbf{b}\right\|_{\infty}\leq\left\|\mathbf{A}\right\|_{\infty}\left\|\mathbf{a}\right\|_{1}\left\|\mathbf{b}\right\|_{1}$ and the fourth inequality

$$\left\|\mathbf{A}\mathbf{B}\mathbf{b}\right\|_{\infty}\leq\left\|\mathbf{A}\mathbf{B}\right\|_{\infty}\left\|\mathbf{b}\right\|_{1}=\max_{i,j}\left|\mathbf{A}_{i\cdot}\mathbf{B}_{\cdot j}\right|\left\|\mathbf{b}\right\|_{1}\leq\left\|\mathbf{A}'\right\|_{c2}\left\|\mathbf{B}\right\|_{c2}\left\|\mathbf{b}\right\|_{1}.$$

The third inequality follows by the Cauchy-Schwarz inequality $\left|\mathbf{a}'\mathbf{A}\mathbf{b}\right|\leq\left\|\mathbf{A}'\mathbf{a}\right\|_{2}\left\|\mathbf{b}\right\|_{2}\leq\left\|\mathbf{A}\right\|_{\mathrm{sp}}\left\|\mathbf{a}\right\|_{2}\left\|\mathbf{b}\right\|_{2}.$ For the symmetric matrix $\mathbf{S}$,

$$\left\|\mathbf{S}\mathbf{b}\right\|_{2}=\sqrt{\mathbf{b}'\mathbf{S}\mathbf{S}\mathbf{b}}\leq\sqrt{\left\|\mathbf{S}\mathbf{S}\right\|_{\infty}}\left\|\mathbf{b}\right\|_{1}\leq\sqrt{\max_{i}\left(\mathbf{S}\mathbf{S}\right)_{ii}}\left\|\mathbf{b}\right\|_{1}=\left\|\mathbf{S}\right\|_{c2}\left\|\mathbf{b}\right\|_{1}$$

where the first inequality follows by (75), and the second inequality and the last equality are due to the symmetry of $\mathbf{Q}$.

The Sherman-Morrison formula gives $\left(\boldsymbol{\Sigma}+\mathbf{a}\mathbf{b}'\right)^{-1}=\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1}\mathbf{a}\mathbf{b}'\boldsymbol{\Sigma}^{-1}/\left(1+\mathbf{a}'\boldsymbol{\Sigma}^{-1}\mathbf{b}\right)$ for any compatible vector $\mathbf{a}$ and $\mathbf{b}$. (79) and (80) follow by setting $\mathbf{b}=\mathbf{a}$ and $\mathbf{b}=-\mathbf{a}$, respectively. ■.

B Additional Numerical Works

B.1 Simulation of the Demonstrative Example

The demonstrative example in Figure 1 is generated from the simple DGP

$$y_t=\sum_{i=1}^{20}w_i x_{it}+\epsilon_t,\quad t=1,\ldots,100$$

The dependent variable y_t is a linear combination of two groups of input variables $\{x_{it}\}_{i=1}^{10}$ and $\{x_{it}\}_{i=11}^{20}$ with group weights $w_i=0.09\cdot 1\left\{1\leq i\leq 10\right\}+0.01\cdot 1\left\{11\leq i\leq 20\right\}$, so that $\sum_{i=1}^{20}w_i=1$. We set $x_{it}\sim$ i.i.d. $N(0,1)$ and $\epsilon_t\sim$ i.i.d. $N(0,0.25)$. We estimate the weights by the ℓ_2-relaxation method under different values of $\tau=0,0.005,0.01,\ldots,0.1$. As the variables within the two groups are symmetric, we report in Figure 1 the empirical bias and variance of $\hat{w}_1$ and $\hat{w}_{11}$ over 1000 replications.

B.2 Simulation of HAR

(The HAR model does not impose the sum-up-to-one restriction. Perhaps we will leave it to another paper.)

HAR is a simple regression model

$$y_t = \sum_{j=1}^{N} \beta_j x_{jt} + \epsilon_t.$$

Given a choice of high frequency, on the right hand side we can come up with as many variables as we wish. We can also introduce dynamic structure to make it more realistic. Corsi (2009)'s idea is to impose a group structure in the predictors. When the imposed group structure is correctly specified, OLS is an oracle estimator. However, it is also possible that the group structure is misspecified.

Even though in the DGP we impose that the coefficients add up as one in the DGP, in estimation of HAR we do not necessarily want to impose need to impose $\mathbf{w}'\mathbf{1}_N = 1$, since strictly speaking HAR is not a forecast combination.

We consider several DGPs.

DGP HAR1: HAR is correctly specified. For example, the day, week and month structure

$$y_t = \frac{\beta_d}{d} \sum_{j=1}^{d} x^{(d)} + \frac{\beta_w}{w} \sum_{j=1}^{w} x^{(w)} + \frac{\beta_m}{m} \sum_{j=1}^{m} x^{(m)} + u_i.$$

In this DGP, if the corresponding regressors for day, week and month are known a prior, OLS will be an oracle estimator that no other method can beat. We hope L2-relaxtion is also competitive.

DGP HAR2: The group information in HAR is misspecified. For example, in the estimation the covariates that share the same coefficients mismatches those in the true DGP. L2-relaxation does not impose the prior knowledge of the group membership so it should be robust. We should find a case of coefficient specification for which OLS the HAR model is not as good as L2-relaxtion.

DGP HAR3: random coefficient model. For example, generate $\beta_j^* = \cos\left(\frac{\pi}{8} j\right) + Unform(0,1)$ for some k and $j = 1, \ldots, 2/N$ and then and $\beta_j^* = 0$ for $j = 0.5N + 1, \ldots, N$. When normalize $\beta_j = \beta_j / \sum_{j=1}^{N} \beta_j$ (to make it sum up to 1). I believe it is possible to come up with a case that L2-relaxtion works better than HAR (depends on the level of mismatch, of course). L2-relaxtion should still be robust, because no matter what is the realized coefficients from the uniform distribution, we can already discretized them into a few groups. (The cosine function also implies deterministic group structure.)

B.3 Empirical Application of HAR

Tian has preliminary results last year.